Processamento e Recuperação de Informação

Web Retrieval and Link Analysis

Departamento de Engenharia Informática
Instituto Superior Técnico

1º Semestre
2018/2019

Bibliography

- Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd edition. Chapter 7.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. Chapters 19 and 21.
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval, 2nd edtion. Chapter 11.

Outline

Traditional IR

Traditional IR systems:

- Worth of a document regarding a query is intrinsic to the document.
- Documents are self-contained units
- Documents are descriptive and truthful

Web IR

The World Wide Web is a shifting universe

- Indefinitely growing and changing
- Non-textual content
- Invisible keywords
- Web spam
- Documents are not self-complete
- Most web queries 2 words long
- Hyperlinked

Many features are included in a web similarity formula

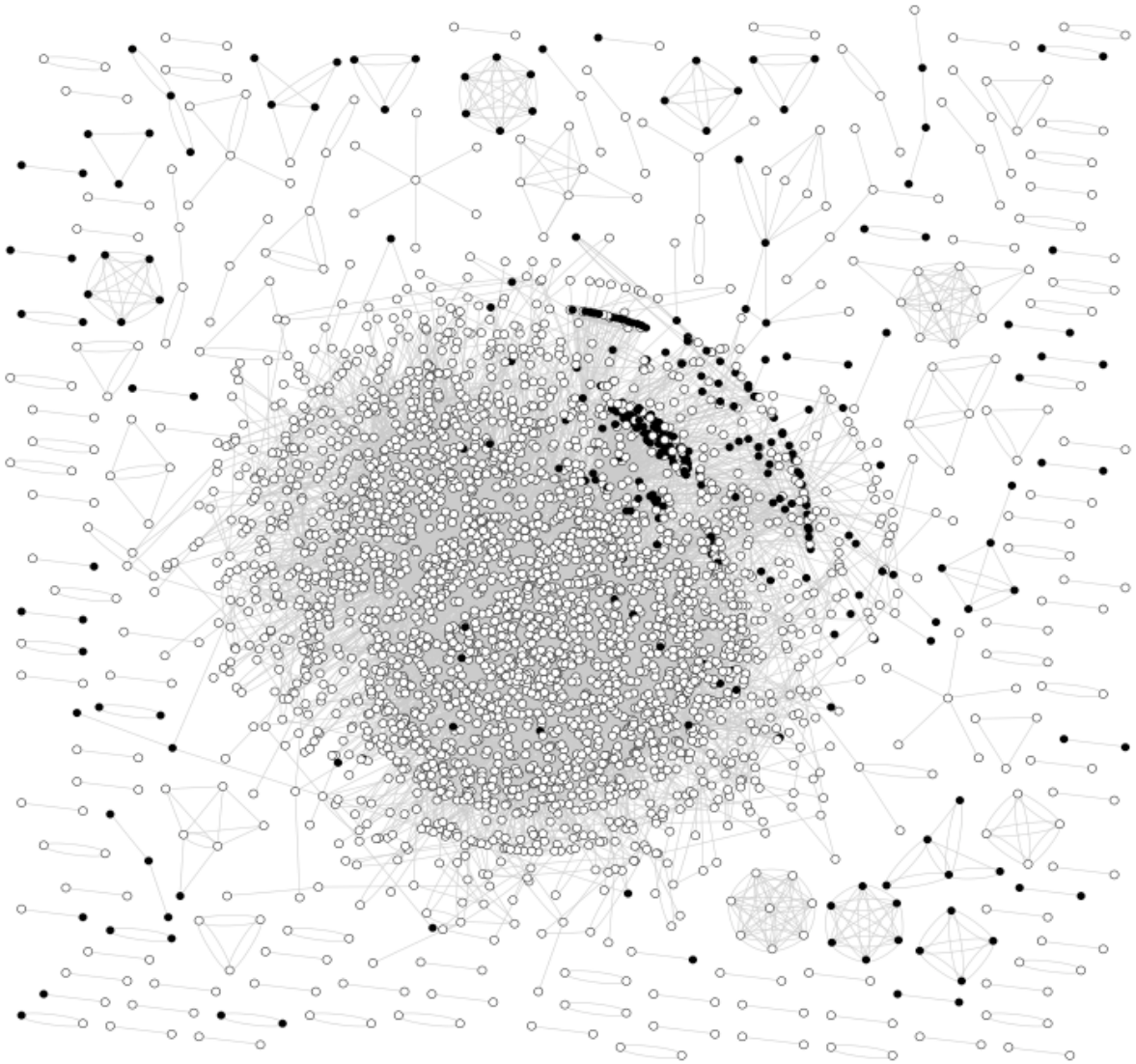
- Ranking functions evaluate the reputation of pages, and different types of content within each page

Outline

The Web as a Graph

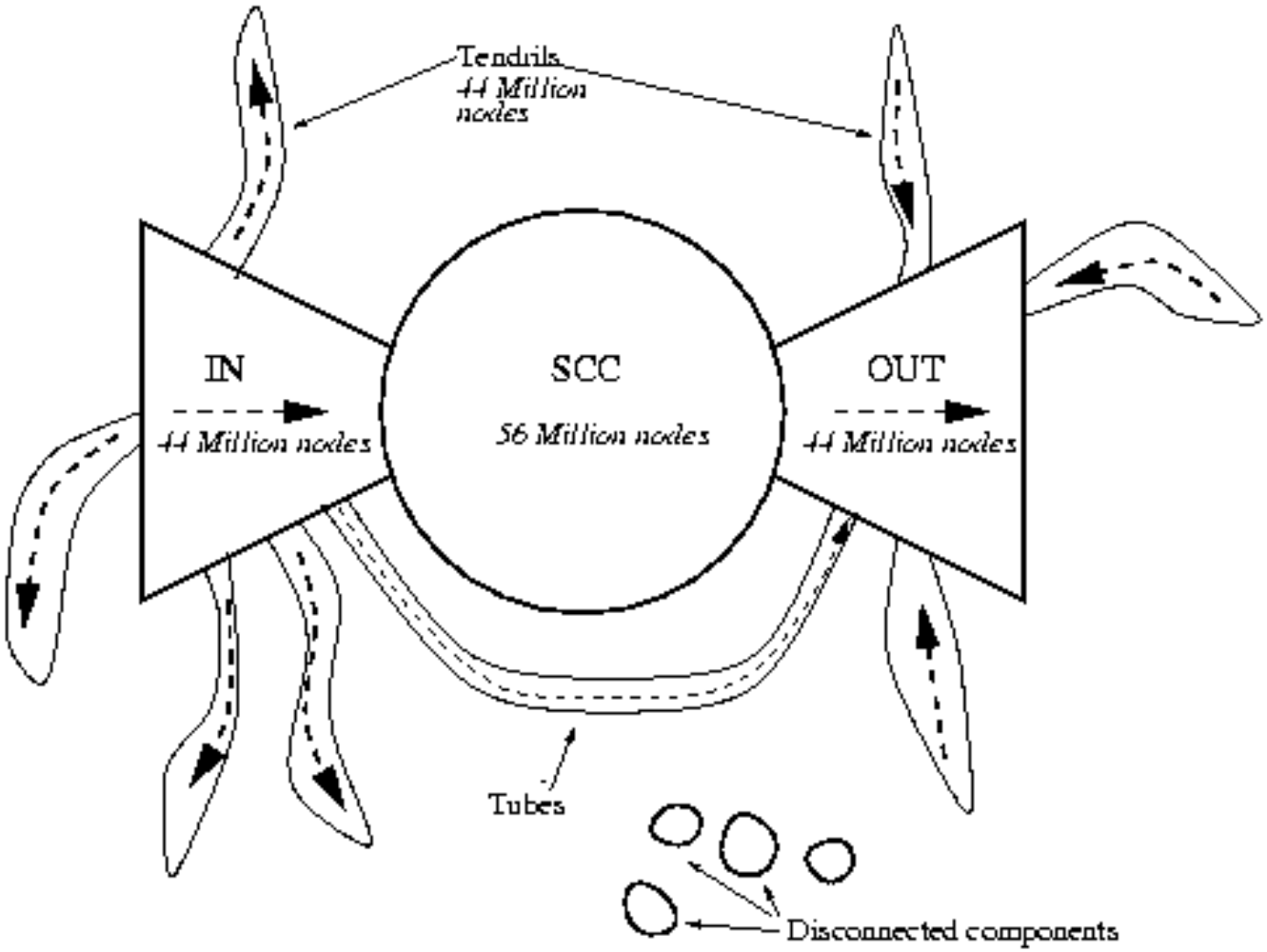
- The Web is an hyperlink graph
  - Evolves organically,
  - No central coordination,
  - Yet shows global and local properties
  - An example of a social network

Graph structure of the Web



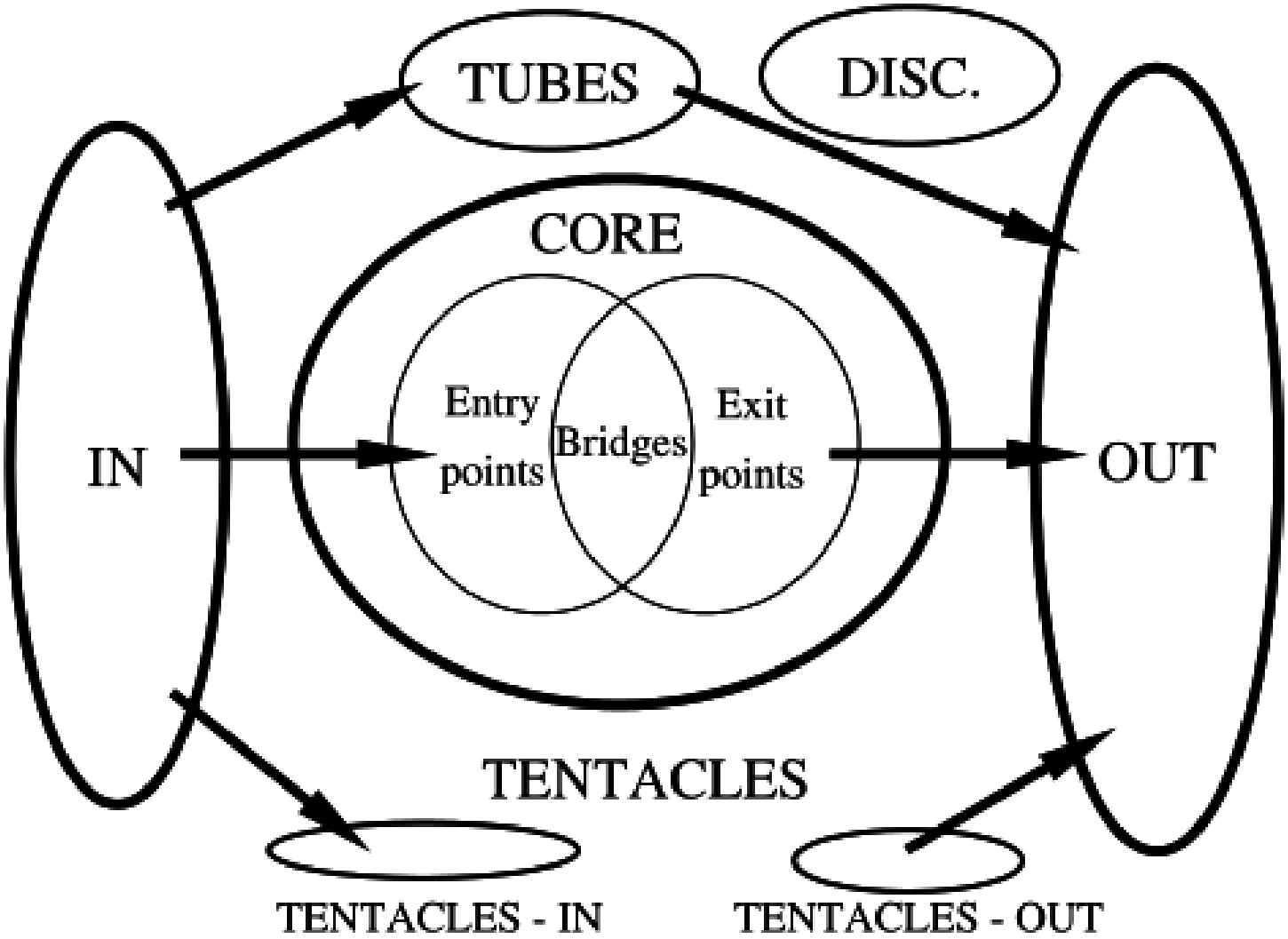
Web Spam detection using the web topology (Castillo et al. 2006)

Graph structure of the Web



The Bow tie model (Broder et al., 2000)

A More Detailed View



A study for link based web page ranking (Baeza-Yates & Castillo, 2001)

A More Recent View (cont.)

- Bridges: sites in CORE that can be reached directly from the IN component and that can reach directly the OUT component
- Entry points: sites in CORE that can be reached directly from the IN component but are not in Bridges
- Exit points: sites in CORE that reach the OUT component directly, but are not in Bridges
- Normal: sites in CORE not belonging to the previously defined sub-components

Outline

Social Network Analysis

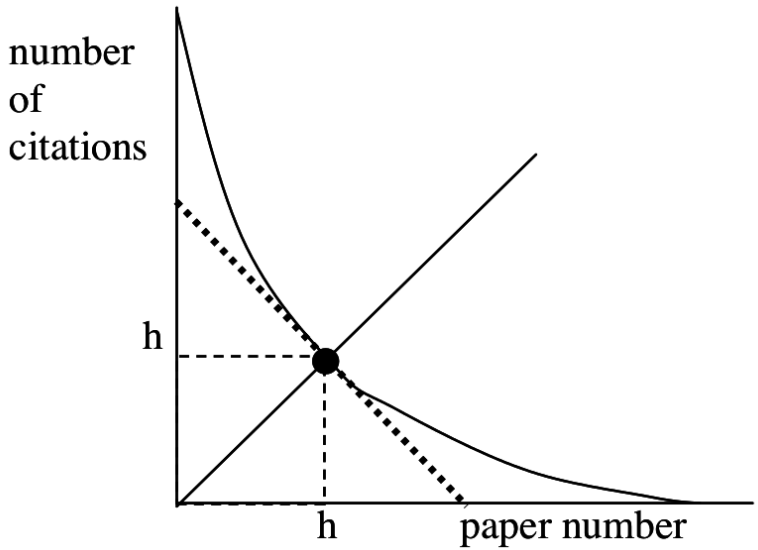
- Social studies based on computing properties related to connectivity and distances in graphs
- Well established, long before the Web
- Example applications:
  - Epidemiology
    - Identifying a few nodes to be removed to significantly increase average path length between pairs of nodes
  - Citation analysis
    - Identifying influential or central papers
    - Identifying influential or central people

Finding influential researchers: H-index

J.E. Hirsch, *An index to quantify an individual's scientific research output*, 2005 (available here)

Definition

Index = h if h papers have at least h citations each, and the remaining papers have no more than h citations each.



Example: `http://www.cs.ucla.edu/~palsberg/h-number.html`

Graph Centrality

- Degree centrality of v
  - Number of edges incident to v
  - Directed graphs: in-degree and out-degree centrality
- Betweenness centrality:

$$C(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

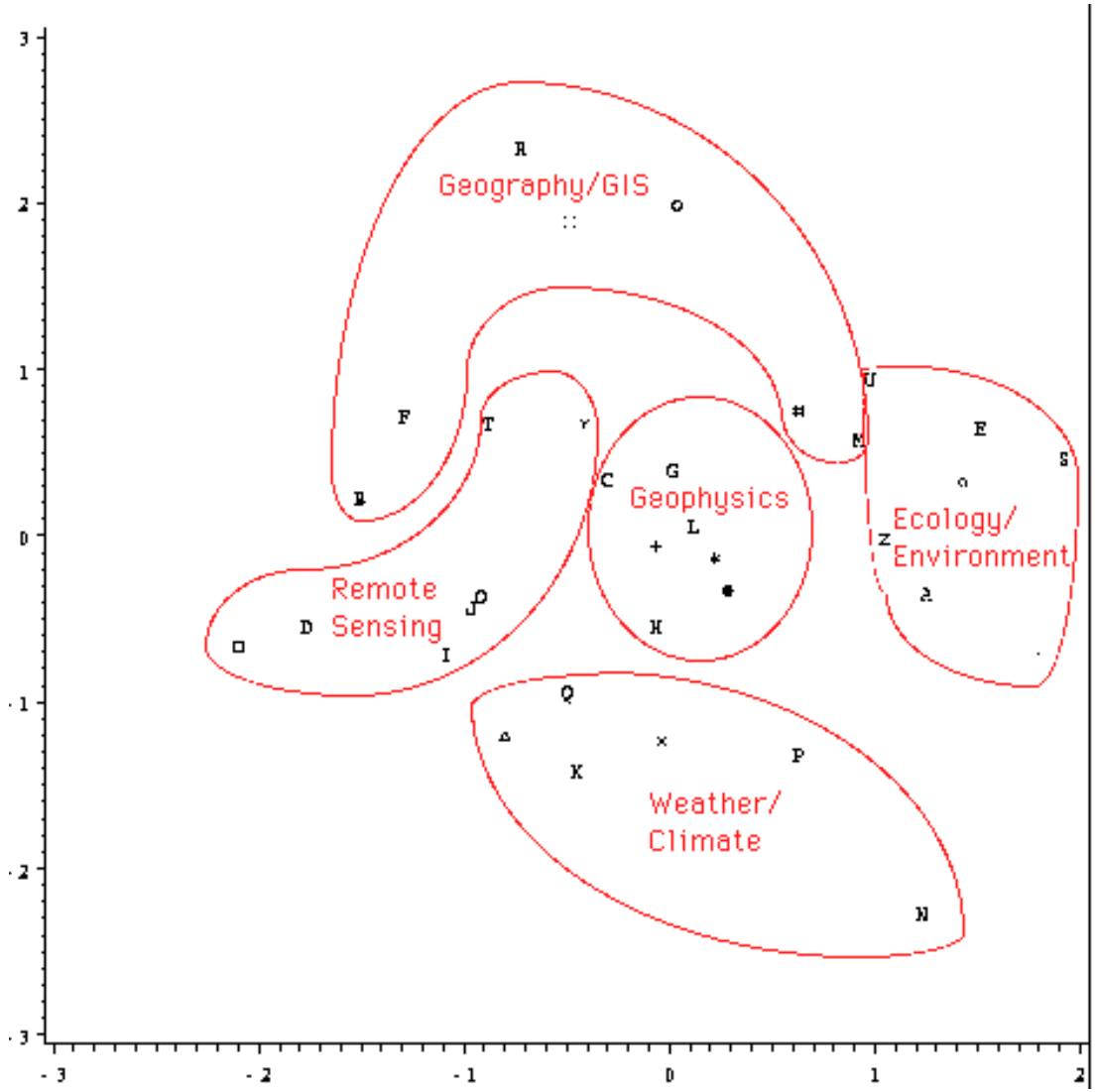
where $\sigma_{st} = \#$ shortest paths from s to t (through v)

Example: http://onearmedman.com/research/swineflu24

Topic similarity using Co-citation

- Documents v and w are said to be co-cited by u if a document u cites documents v and w
- If E is the document citation matrix
  - $E^T E$ is the co-citation index matrix
  - Indicator of relatedness between every v and w
- Example use: clustering
  - Using above pair-wise relatedness measure in a clustering algorithm

Clustering with Co-citation



Social structure of Web communities (source: Ray R. Larson).

Outline

The Web as a Graph

- Hypermedia is a graph of documents
- We can apply social network theory
  - Extensive research applying graph notions
  - Centrality and prestige
  - Co-citation
- Application: link analysis

Link Analysis

Three levels of analysis:

- Macroscopic: related to the structure of the Web at large
  - E.g. the bow tie structure analysis
- Mesoscopic: related to the properties of areas or regions of the Web
  - E.g. link-based ranking
- Microscopic: related to the statistical properties of links and individual nodes
  - E.g. link properties

Using Link Analysis

Main applications:

- Prioritize crawling
- Identify sub-structures on the Web graph, such as communities
- Infer relevance

Link-based Ranking Strategies

Goal: Leverage linkage information to address the *abundance problems* inherent in broad queries

Two pioneering algorithms:

- **PageRank**: Measure of prestige for every page on web
- **HITS**: Identify *hubs* and *authorities* in a query result

Link Model

- Each page is a node without any textual properties
- Each hyperlink is an edge connecting two nodes (possibly with an edge weight)
- Some preprocessing procedure outside the scope of the algorithm may be used to choose what sub-graph of the Web to analyze

Outline

Overview

- Pre-computes a rank-vector
  - Provides a-priori (offline) importance estimates for all pages on Web (i.e., probability distribution over pages)
  - Independent of the search query
- Prestige $\approx$ In-degree
- But not all votes are worth the same
- Prestige of a page is proportional to the sum of the prestige of citing pages
- PageRank is part of the ranking strategy adopted by Google
  - At query time: prestige scores used in conjunction with query-specific IR scores

PageRank Algorithm

The algorithm:

- E is the *adjacency matrix* of the Web

$$E[u,v] = \begin{cases} 1 \text{ iff there is a link from } u \text{ to } v \\ 0 \text{ otherwise} \end{cases}$$

- The out-degree of node u is given by

$$N_u = \sum_v E[u,v]$$

- Start with an initial prestige vector $p_0[u]$
- Compute

$$p_{i+1}[v] = \sum_{(u,v)\in E} \frac{p_i[u]}{N_u}$$

Main Features

- PageRank simulates a user navigating randomly on the Web
- At infinity, the probability of finding the user at any given page becomes stationary
- This process can be modeled by a Markov chain
  - stationary probability of being at each page can be computed
- This probability is a property of the graph
  - referred to as PageRank in the context of the Web

Convergence Conditions

- Convergence to
  - stationary distribution of the normalized adjacency matrix L
  - PageRank vector p is principal eigenvector of L
- Convergence criteria
  - L is irreducible
    - there is a directed path from every node to every other node
  - L is aperiodic
    - for every node, there is no integer $k > 1$ that divides the length of every cycle that goes through the node
    - the reverse—periodic—means that a node is visited always after a regular nk number of steps ($n = 1, 2, 3, \ldots$)

Problems of Convergence

- Web graph is not strongly connected
  - Only a fourth of the graph is!
- Web graph is not aperiodic
  - There can be many periodic nodes in the Web graph
- Rank-sinks
  - Pages without out-links
  - Directed cyclic paths

A simple fix

- Two way choice at each node:
  - With a certain probability d $(0.1 < d < 0.2)$, the surfer jumps to a random page on the Web
  - With probability $1 - d$ the surfer decides to choose, uniformly at random, an out-neighbor

$$p_{i+1}[v] = \frac{d}{N} + (1 - d) \sum_{(u,v) \in E} \frac{p_i[u]}{N_u}$$

Dangling Links

- Dangling Links are links to pages with no outgoing links
- These pages receive PageRank but do not share it with other pages
  - In a sense, they are *wasting* their PageRank
- Ways to deal with it:
  - Remove the links from the computation and add them in final iteration (original proposal)
  - Consider a virtual link from the linked page to all other pages

PageRank architecture at Google

- Ranking of pages more important than exact values of p
- Convergence of page ranks in 52 iterations for a crawl with 322 million links.
- Pre-compute and store the PageRank of each page.
  - PageRank independent of any query or textual content.
- Ranking scheme combines PageRank with textual match
  - Unpublished *learning-to-rank* approach
  - Many empirical parameters, human effort and regression testing.

Other Applications and Extensions

- PageRank also used in other IR/IE applications (e.g., text summarization, keyword extraction, etc.)
- Many extensions proposed over the years
  - PageRank with edge weights:

$$p_{i+1}[v] = \frac{d}{N} + (1-d) \sum_{(u,v)\in E} \frac{p_i[u] \times w(u,v)}{\sum_{(u,v')\in E} w(u,v')}$$

  - Personalized PageRank:

$$p_{i+1}[v] = \frac{d \times w(v)}{\sum'_v w(v)} + (1-d) \sum_{(u,v)\in E} \frac{p_i[u]}{N_u}$$

  - Topic-sensitive PageRank, etc.

Outline

HITS: Hypertext Induced Topic Selection

- Relies on query-time processing
  - To select base set V_q of links for query q constructed by
    - selecting a sub-graph R from the Web (root set) relevant to the query
    - selecting any node u which neighbors any $r \in R$ via an inbound or outbound edge (expanded set)
  - To deduce hubs and authorities that exist in a sub-graph of the Web
- Every page u has two distinct measures of merit,
  - its hub score $h[u]$
  - its authority score $a[u]$
- Recursive quantitative definitions of hub and authority scores

Hubs and Authorities

Hub

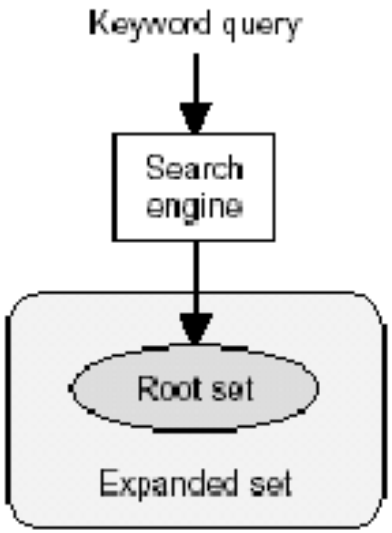
A page is a good hub if it contains links to many good authority pages

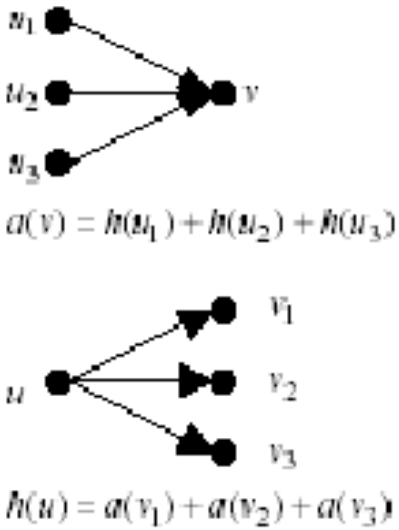
Authority

A page is a good authority if it is pointed to by many good hubs

- Authority pages provide good content.
- Hub pages provide links to the pages with good content.

The HITS Algorithm





$\vec{a} \leftarrow (1, \ldots, 1)^T$, $\vec{h} \leftarrow (1, \ldots, 1)^T$
while $\vec{h}$ and $\vec{a}$ change 'significantly' do
  $\vec{h} \leftarrow E\vec{a}$
  $\ell_h \leftarrow \|\vec{h}\|_1$
  $\vec{h} \leftarrow \vec{h}/\ell_h$
  $\vec{a} \leftarrow E^T\vec{h}_0 = E^T E\vec{a}_0$
  $\ell_a \leftarrow \|\vec{a}\|_1$
  $\vec{a} \leftarrow \vec{a}/\ell_a$
end while

https://en.wikipedia.org/wiki/HITS_algorithm

Some Issues

- Does not work with non-existent, repeated, or automatically generated links
  - Solution: weigh each link based on surrounding content
- Topic diffusion
  - The result set might include pages that are not directly related to the query
  - One solution: associate a score with content of each page
  - This score is then combined with the link weight
  - Experiments show that recall and precision for first ten results increase significantly

PageRank vs. HITS

- PageRank advantage over HITS
  - Query-time cost is low
    - HITS computes an eigenvector for every query
  - Less susceptible to localized link-spam
- HITS advantage over PageRank
  - HITS ranking is sensitive to query
  - HITS has notion of hubs and authorities

Outline

Web Spamming

Activity of deliberately misleading a search engine by a website owner.
Deceivers try to understand how a ranking function computes, by changing the ranking of a page without changing its user-perceived value.

SEO - Search Engine Optimization:

A business activity that sometimes is legitimate, but often is not perceived as ethical.

Content Spamming

Attempt to affect the content-based ranking features

Places where to add spam terms:

- Title
- Meta-tags
- Body
- Anchor text
- URL

Techniques

- repeat some important terms

  the picture quality of the camera is of amazing quality

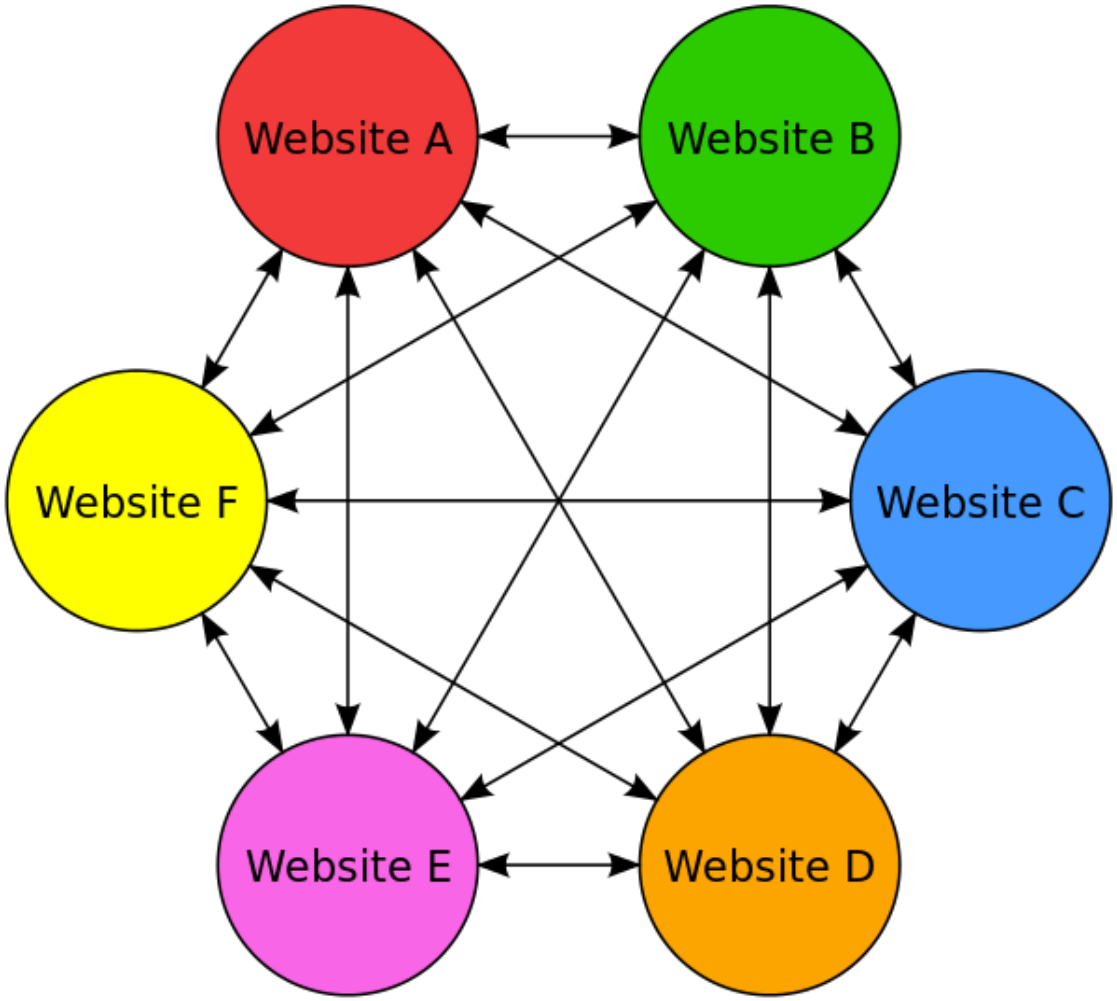
- dumping of many unrelated terms

  Ed Sheeran Dua Lipa Sam Smith Eminem Pink Camila

Link Spamming

- out-link spamming
  - easy: pick popular websites from directories
- in-link spamming
  - Create a honey pot
  - Add links to web directories
  - Post links to user-generated content sites
  - Participate in a link exchange
  - Create a spam farm

Link Spamming



http://en.wikipedia.org/wiki/Link_farm

Hiding Techniques

Content hiding: pick background white and font color also white

Cloaking: serve one page to normal clients and another to search engines

Redirection: redirect browser to another page (user sees one, search engine will crawl both)

URL Redirection

https://en.wikipedia.org/wiki/URL_redirection

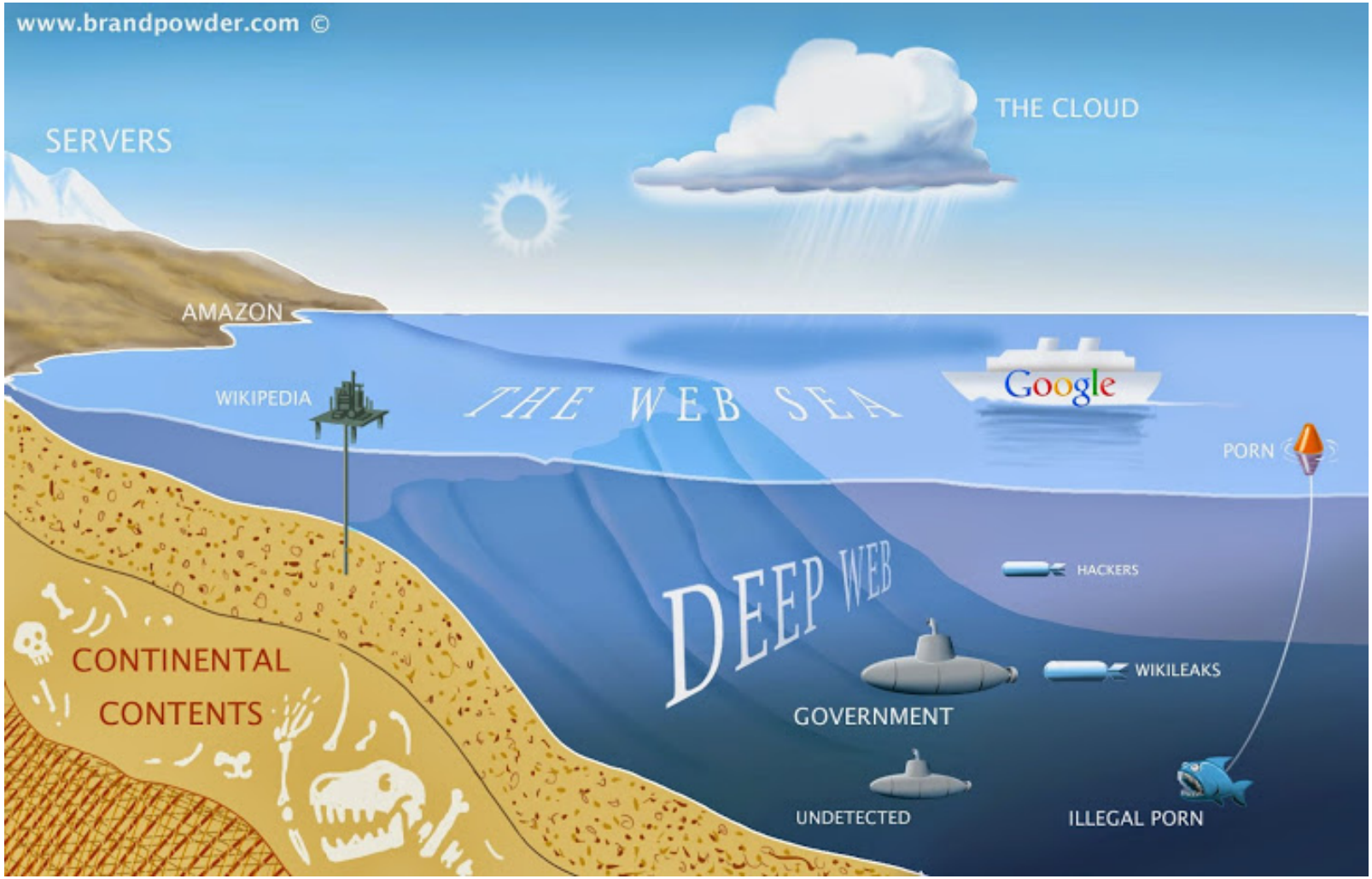
Combating Spam

- Give higher weight to anchor text
- PageRank - assign authority to pages based on number and importance of links
- TrustRank - the good guys and the bad guys cluster together
- Learn from language features common in spam (longer titles, longer words, ...)
- Partition pages in blocks and compute PageRank on a block basis (instead of assigning a single PR value to each page), to defeat honeycombs and link exchanges
- ...an on-going process...

Click Farming

https://en.wikipedia.org/wiki/Click_farm

The Deep Web



http://thehackernews.com/2012/05/what-is-deep-web-first-trip-into-abyss.html

Questions?