Link Analysis

from Bing Liu.

"Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer

and other material.

Contents

Introduction

- **Early search engines mainly compare content similarity of the query and the indexed pages, i.e.,**
  - **they use information retrieval methods, cosine, TF-IDF, ...**
- **From 1996, it became clear that content similarity alone was no longer sufficient.**
  - **The number of pages grew rapidly in the mid-late 1990's.**
    - **Try the query "Barack Obama". Google estimates about 140,000,000 relevant pages.**
    - **How to choose only 30-40 pages and rank them suitably to present to the user?**
  - **Content similarity is easily spammed.**
    - **A page owner can repeat some words (TF component of ranking) and add many related words to boost the rankings of his pages and/or to make the pages relevant to a large number of queries.**

Introduction (cont …)

- **Starting around 1996, researchers began to work on the problem. They resort to hyperlinks.**
  - **In Feb, 1997, Yanhong Li (Scotch Plains, NJ) filed a hyperlink based search patent. The method uses words in anchor text of hyperlinks.**
- **Web pages on the other hand are connected through hyperlinks, which carry important information.**
  - **Some hyperlinks: organize information at the same site.**
  - **Other hyperlinks: point to pages from other Web sites. Such out-going hyperlinks often indicate an implicit conveyance of authority to the pages being pointed to.**
- **Those pages that are pointed to by many other pages are likely to contain authoritative information.**

Introduction (cont …)

- **During 1997-1998, two most influential hyperlink based search algorithms, namely PageRank and HITS, were reported.**
- **Both algorithms are related to social networks. They exploit the hyperlinks of the Web to rank pages according to their levels of "prestige" or "authority".**
  - **HITS: Jon Kleinberg (Cornel University), at *Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, January 1998**
  - **PageRank: Sergey Brin and Larry Page, PhD students from Stanford University, at *Seventh International World Wide Web Conference* (WWW7) in April, 1998.**
- **PageRank powered the first releases of the Google search engine.**

Introduction (cont …)

- **Apart from search ranking, hyperlinks are also useful for finding Web communities.**
  - **A Web community is a cluster of densely linked pages representing a group of people with a special interest.**
- **Beyond explicit hyperlinks on the Web, implicit/explicit links in other contexts are useful too, e.g.,**
  - **for discovering communities of named entities (e.g., people and organizations) in free text documents, and**
  - **for analyzing social phenomena in emails.**

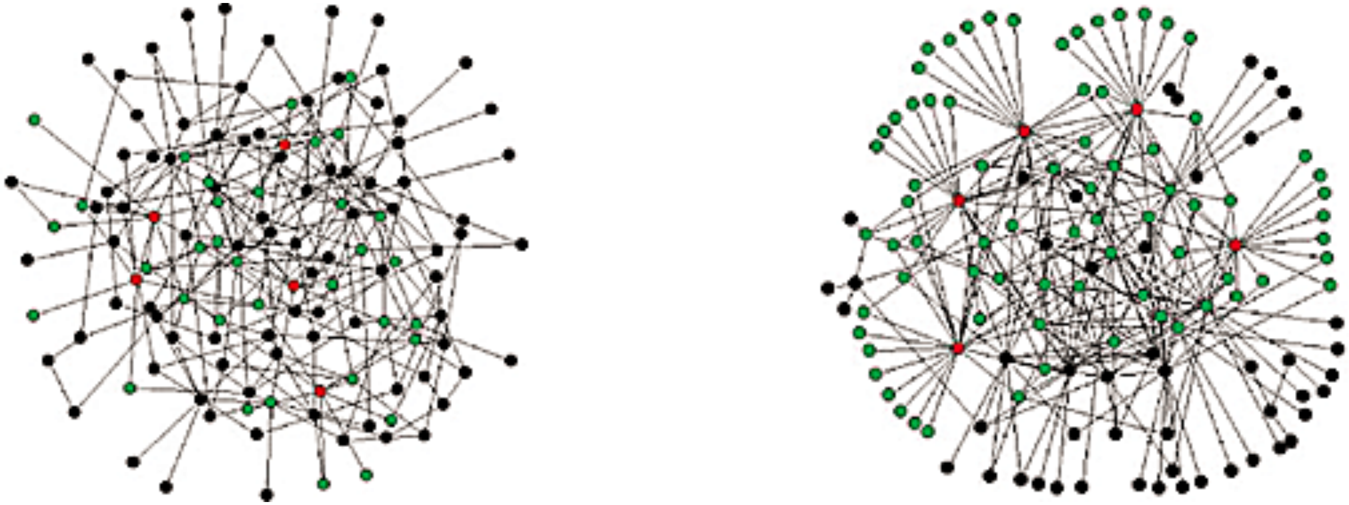
Contents

Network properties

- **The Web graph shares many properties with other graphs/networks**
  - **networks of the co-authors, phone calls, emails, citations, software classes, neurons, protein interactions**

Network properties

- **Using a Web crawler, Barabasi-Albert in 1999 mapped the connectedness of the Web, and coined the term scale free**



random *scale free*

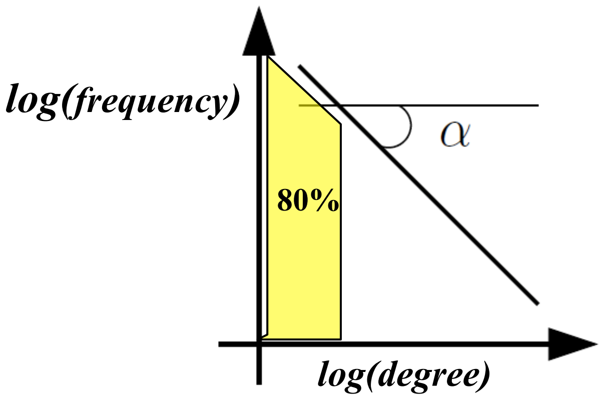
- **A network is scale-free if its degree distribution (**the probability $P(k_i)$ that a node selected uniformly at random has k_i degree**) follows a power law**

Albert-László Barabási and Réka Albert, **"Emergence of scaling in random networks"**, Science, 286:509-512, October 15, 1999

Network properties

- **Scale-free networks**
  - **Power law of the graph degree**
  - **Pareto principle 80:20 : roughly 80% of the effects come from 20% of the causes**
    "80:20 rule"

$$p(k) \approx c \cdot k^{-\alpha}$$



  - **The ratio of very connected nodes to the number of nodes in the rest of the network remains constant as the network changes in size**
  - **Scale-free networks may show almost no degradation as random nodes fail**
    - **If failures occur at random, since the vast majority of nodes are those with small degree, the likelihood that a hub would be affected is almost negligible. In addition, we have the clustering property of these networks (see the following slides)**

Network properties

- **Scale-free networks**
  - A power law looks the same, no matter what scale we look at it
    - The *shape of the distribution* is unchanged, except for a multiplicative constant, i.e. scale-free
    - e.g.: it looks the same from 2 to 50 or scaled up (hundred fold) from 200 to 5000
  - Given a probability distribution $p(x)$, it is scale-free if exists $g(b)$ such that $p(bx) = g(b)\, p(x)$ for each b and x
  - Given a power-law: $p(x) = c\, x^{-\alpha}$
    - $p(bx) = c\, (bx)^{-\alpha} = b^{-\alpha}\, c\, x^{-\alpha}$ <= $g(b) = b^{-\alpha}$

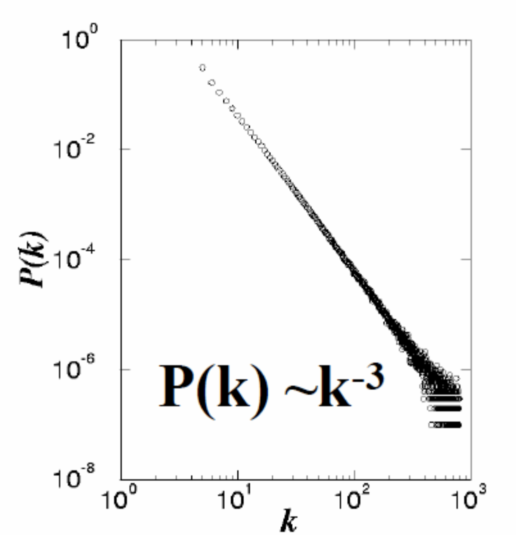
Network properties

- **Scale-free networks arise from preferential attachment process**
  - **basically it is a pattern of network growth in which the rich (nodes) get richer (i.e., more edges)**
  - **Think of citation networks, like the World Wide Web. You are more likely to cite (link to) pages that already have a lot of links, and hence there is a positive feedback loop**
  - **Demo: http://ccl.northwestern.edu/netlogo/models/PreferentialAttachment**

Network properties

Preferential attachment process

- (1) Networks continuously expand by the addition of new nodes
  - WWW : addition of new documents
- (2) New nodes prefer to link to highly connected nodes
  - WWW : linking to well known sites



GROWTH:

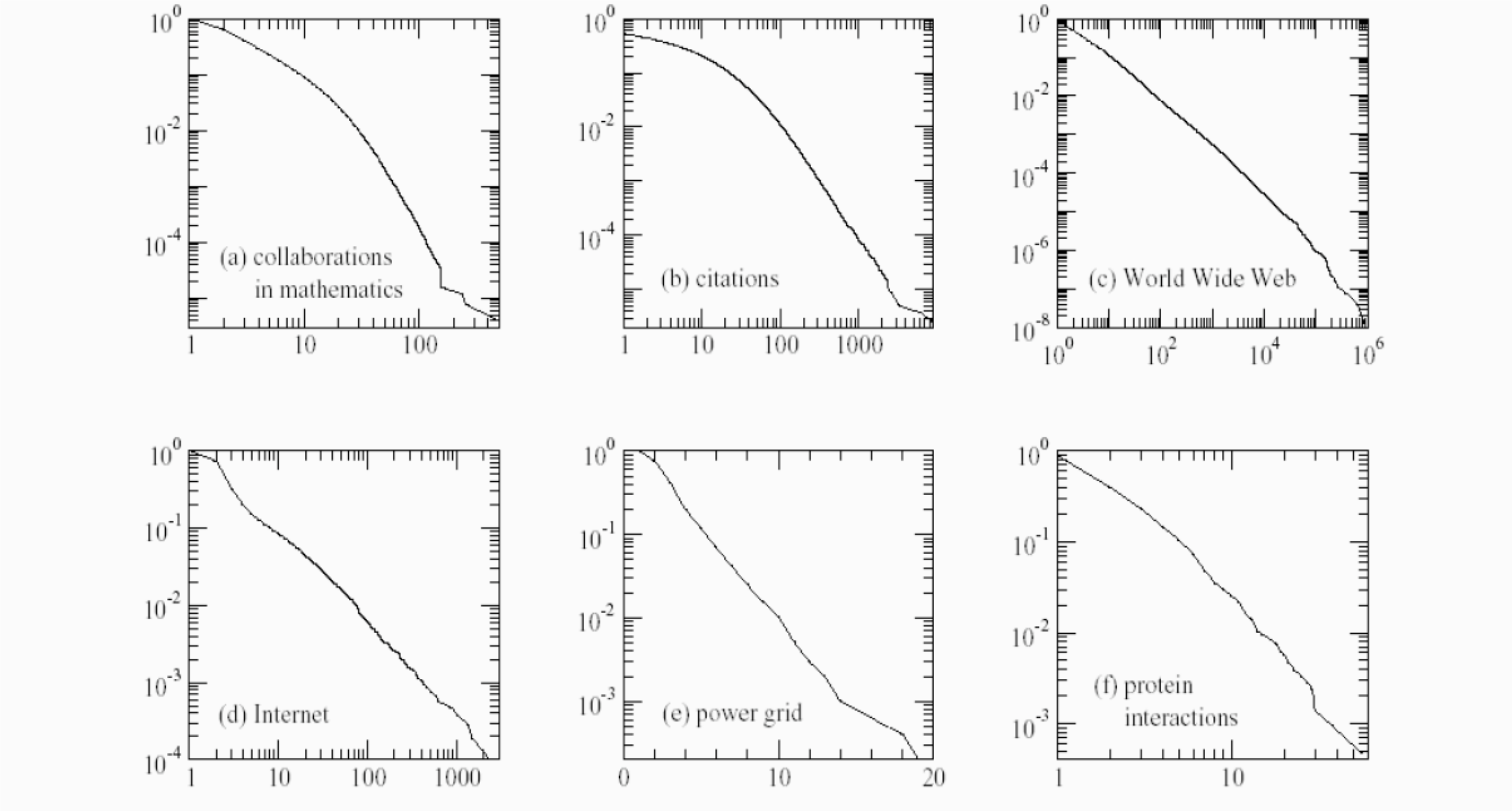
add a new node with m links

PREFERENTIAL ATTACHMENT:

the probability that the new node connects to $node_i$ is proportional to the number of existing links k_i that $node_i$ already has

$$P(linking\_to\_node_i) = \frac{k_i}{\sum_j k_j}$$

Examples of scale free networks



Network properties

- **Scale free networks**
  - have also a high value of the so-called clustering coefficient C
    - "My friends are friends too"
  - Given a vertex v, let k_v be the number of its neighbors
    - $k_v (k_v-1)$: max allowable number of (directed) links among all the k_v neighbors
    - C_v is the fraction of the allowable links that actually exist
  - C is the average over all C_v

- **Clustering coefficient C of the scale-free network is about five times higher than the coefficient of a random graph**
  - Factor slowly increases with the number of nodes
  - This increases the resilience to node failure

Network properties

- **Small world networks**
  - **Let d be the mean of the shortest paths**
    - **d is *small* w.r.t. the number of nodes N**
  - ***Milgram* experiment(1967)**
    - **only six degrees of separation between any two people in the world ("*know*" relationship)**
- **The Web network is also "small world"**
  - **note that also a random networks (not only a scale-free one) can have a short average path length d**
  - **the average of d over all pairs of vertices follows:**

    $d = 0.35+2.06 \log(N)$
  - **using N = 8×10^8, we have $d_{www} = 18.59$, i.e., two randomly chosen documents on the web are on average 19 clicks away from each other**

Réka Albert, Hawoong Jeong, and Albert-László Barabási (1999). **"The Diameter of the WWW"**. Nature 401: 130-131.

Contents

Social network analysis

- Social network
  - the study of social entities (e.g. people in an organization, called actors), and their interactions and relationships
- The interactions and relationships can be represented with a network or graph
  - each vertex (or node) represents an actor, and
  - each link represents a relationship.
- We can study the properties of the network structure
  - role, position and prestige of each social actor.
  - finding various kinds of sub-graphs, e.g., communities formed by groups of actors.

Social network and the Web

- **Social network analysis is useful for the Web**
  - **the Web is essentially a virtual society, and thus a virtual social network**
  - **each page: a social actor**
  - **each hyperlink: a relationship.**
- **Many results from social network can be adapted and extended, and used in the Web context.**
- **We study two types of social network analysis, centrality and prestige, which are closely related to hyperlink analysis and search on the Web.**

Centrality

- **Important or prominent actors are those that are linked or involved with other actors extensively.**
- **A person with extensive contacts (links) or communications with many other people in the organization is considered more important than a person with relatively fewer contacts.**
- **The links can also be called ties.**
- **A central actor is one involved in many ties.**

Degree Centrality

- **Concept based on the direct connections, only out-links in directed graphs**

- **Undirected graph:**
  - **normalized node degree, where *d(i)* is the degree of node *i* and *n* is the number of nodes**

$$C_D(i) = \frac{d(i)}{n-1}$$

- **Directed graph:**
  - only **out-links**

$$C'_D(i) = \frac{d_o(i)}{n-1}$$

Closeness Centrality

- **Concept based on the distance between node pairs**
  - ***d(i,j)* is the shortest distance between two nodes**
  - **we are supposing that the graph is connected, and thus i.e., there is a path from any point to any other point in the graph:** $d(i,j) \geq 0$

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

- **Ranges between 0 and 1**
- **Directed graph**
  - **consider edge direction in computing the distances**

Betweenness Centrality

- **Suppose two non-adjacent actors *j* and *k* want to interact**
  - **if actor *i* is on the path between *j* and *k*, then *i* may have some control over the interactions between *j* and *k*.**
- **Betweenness measures this control of *i* over other pairs of actors.**
  - **Thus, if *i* is on the paths of many such interactions, then *i* is an important actor.**
  - **In this case it plays a 'broker' role in the network.**
  - **The good news is that *i* plays a powerful role in the network, the bad news is that it is a single point of failure.**

Betweenness Centrality (cont …)

- **Undirected graph:**
  - **Let p_{jk} be the number of shortest paths between actor j and actor k, where $i \neq j$ and $i \neq k$**
  - **More than one shortest path may exist**

- **$p_{jk}(i)$ is the number of shortest paths that pass i**

- **The betweenness of an actor i is defined as the sum of the various $p_{jk}(i)$ normalized by the total number of shortest paths**

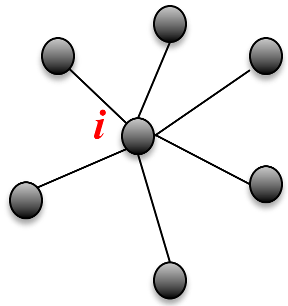
$$\sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$$

Betweenness Centrality (cont …)

- The coefficient can be further normalized by dividing by *(n-1)(n-2)/2*, which is the maximum quantity for the above formula (when all the shortest paths from *j* to *k* pass *i*)

$$C_B(i) = \frac{2\sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}$$

$$C_B(i) = \frac{2(6 \cdot 5/2)}{6 \cdot 5} = 1$$



- For a directed graph we need to consider that $p_{jk} \neq p_{kj}$

Prestige

- **Prestige is a more refined measure of prominence of an actor than centrality.**
  - **We have to distinguish between: ties sent (out-links) and ties received (in-links).**
- **A prestigious actor is one who is object of extensive ties as a recipient.**
  - **To compute the prestige: we only use in-links.**
- **Difference between centrality and prestige:**
  - **centrality focuses on out-links (or undirected graphs)**
  - **prestige focuses on in-links.**
- **We study three prestige measures. Rank prestige forms the basis of most Web page link analysis algorithms, including PageRank and HITS**

Degree prestige

- $d_I(i)$ is the in-degree of node I
  - the degree is normalized to obtain a measure between 0 and 1

$$P_D(i) = \frac{d_I(i)}{n-1}$$

- A node is thus prestigious if it receives many in-links or nominations

Proximity prestige

- The Degree Prestige of actor i only considers the adjacent actors
- The proximity prestige generalizes it
  - We consider every actor j that can reach i
- Let I_i be the set of actors that can reach actor i
- The proximity is defined in terms of closeness or distance of the other actors in I_i to i
- Let $d(j, i)$ denote the distance (shortest path) from actor j to actor i
  - This measure is directly proportional to the distance of all the actors in I_i from i

$$\sum_{j \in I_i} \frac{d(j,i)}{|I_i|}$$

Proximity prestige (cont …)

- **The Proximity Prestige index is the following**
  - **the numerator is the fraction of the actors that can reach i**
  - **the denominator is $\geq I_i$**
  - **$P_P(i)$ thus ranges between 0 and 1**
    - **it is 1 when all the n-1 nodes are directly connected to i (and thus $|I_i| = n$-1)**

$$P_p(i) = \frac{\dfrac{|I_i|}{n-1}}{\displaystyle\sum_{j \in I_i} \frac{d(j,i)}{|I_i|}}$$

Rank prestige

- **In the previous two prestige measures, an important factor is not considered:**
  - **the prominence of individual actors who do the "voting"**
- **In the real world, a person i chosen by an important person is more prestigious than one chosen by a less important person.**
- **If one's circle of influence is full of prestigious actors, then one's own prestige is also high.**
  - **Thus one's prestige is affected by the ranks or statuses of the involved actors.**

Rank prestige (cont …)

- Based on this intuition, the **rank prestige** $P_R(i)$ is define as a **linear combination** of the prestige ranks of the actors whose links point to i:

$$P_R(i) = \sum_{j=1}^{n} A_{ji} \cdot P_R(j)$$

  where $A_{ji}=1$ if j points to i, 0 otherwise
- Let P be the column vector of all the rank prestige:

$$P = (P_R(1), P_R(2), \ldots, P_R(n))^T$$

Rank prestige (cont …)

- **Then we can write:**

$$P = A^T P$$

where A is the 0/1 incidence matrix, where $A(i,j)=1$ if node i point to j, 0 otherwise

- **This is the characteristic equation used to find the eigenvector system of matrix A^T**
  - **P is an eigenvector of matrix A^T with eigenvalue equal 1**
- **The PageRank algorithm computes the Rank Prestige**

Contents

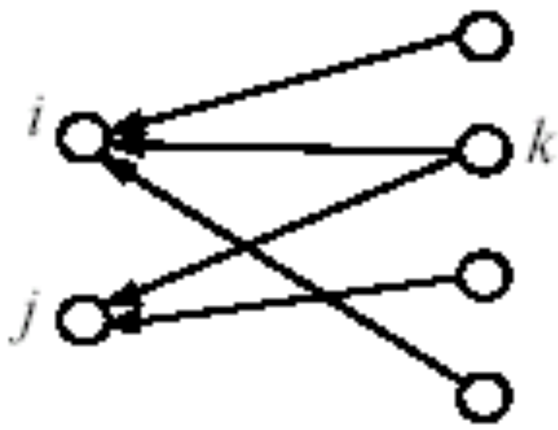
- Introduction
- Network properties
- Social network analysis
- Co-citation and bibliographic coupling
- PageRank
- HITS
- Summary

Co-citation and Bibliographic Coupling

- Another area of research concerned with links is citation analysis of scholarly publications
- When a paper cites another paper, a relationship is established between the publications.
  - Citation analysis uses these relationships (links) to perform various types of analysis.
- We discuss two types of citation analysis, co-citation and bibliographic coupling. The HITS algorithm is related to these two types of analysis.

Co-citation

- If papers i and j are both cited by paper k, then they may be related in some sense to one another.
- The more papers they are cited by, the stronger their relationship is



Co-citation

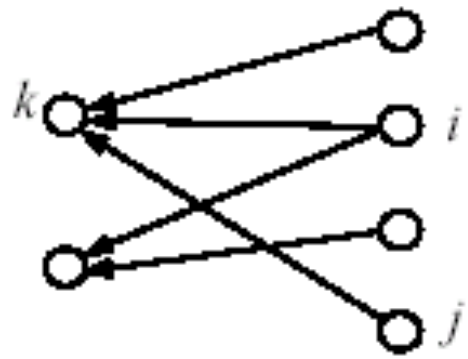
- **Let L be the citation matrix. Each cell of the matrix is defined as follows:**
  - $L_{ij} = 1$ **if paper i cites paper j, and 0 otherwise.**
- **Co-citation (denoted by C_{ij}) is a similarity measure defined as the number of papers that co-cite i and j**

$$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$$

- **C_{ii} is naturally the number of papers that cite i**
- **A square matrix C can be formed with C_{ij}, and it is called the co-citation matrix**

Bibliographic coupling

- Bibliographic coupling operates on a similar principle.
- Bibliographic coupling links papers that cite the same articles
  - if papers i and j both cite paper k, they may be related.
- The more papers cite both papers i and j, the stronger their similarity is.



Bibliographic coupling (cont …)

- **Bibliographic coupling** (denoted by B_{ij}) is a similarity measure defined as the number of papers that are cited by both papers i and j

$$B_{ij} = \sum_{k=1}^{n} L_{ik} L_{jk}$$

- B_{ij} is also symmetric and can be used to measure the similarity of two papers in clustering

Contents

- **Introduction**
- **Graph properties**
- **Social network analysis**
- **Co-citation and bibliographic coupling**
- **PageRank**
- **HITS**
- **Summary**

PageRank

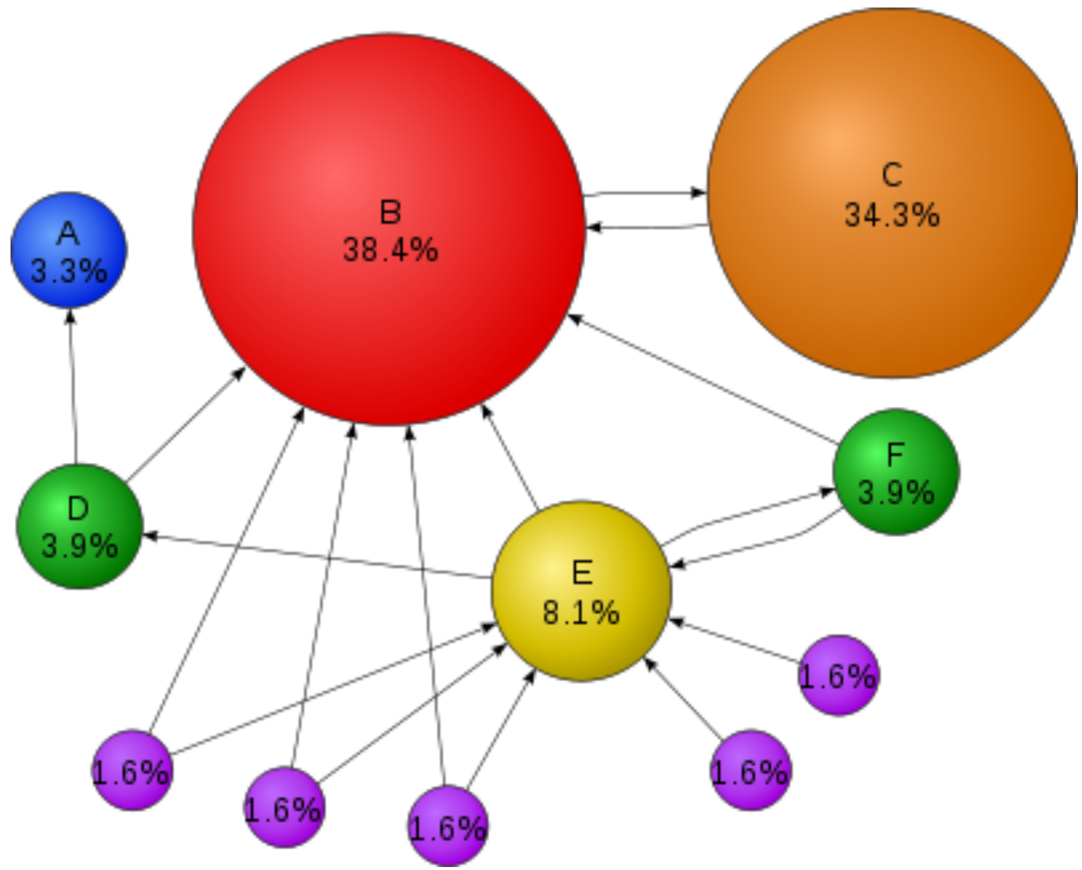
- The year 1997/1998 were eventful years for Web link analysis models. Both the PageRank and HITS algorithms were reported in that year.
- The connections between PageRank and HITS are quite striking.
- Since that eventful year, PageRank has emerged as the dominant link analysis model
  - due to its query-independence
  - its ability to combat spamming
  - Google's huge business success.

Brin, S. and Page, L. (1998) ***The Anatomy of a Large-Scale Hypertextual Web Search Engine***. *In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.*

J. Kleinberg. ***Authoritative sources in a hyperlinked environment***. *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.*

PageRank

- **PageRank is a link analysis algorithm, named after Brin & Page [1], and used by the Google Internet search engine, which assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set [wikipedia]**



[1] Brin, S. and Page, L. (1998) ***The Anatomy of a Large-Scale Hypertextual Web Search Engine***. *In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia*

PageRank: the intuitive idea

- **PageRank relies on the democratic nature of the Web** by using its vast link structure as an indicator of an individual page's value or quality.
  - PageRank interprets a hyperlink from page x to page y as a vote, by page x, for page y
- However, PageRank looks at more than the sheer number of votes; it also analyzes the page that casts the vote.
  - Votes casted by "important" pages weigh more heavily and help to make other pages more "important"
- This is exactly the idea of **rank prestige** in social network.

More specifically

- **A hyperlink from a page to another page is an implicit conveyance of authority to the target page.**
  - **The more in-links that a page i receives, the more prestige the page i has.**
- **Pages that point to page i also have their own prestige scores.**
  - **A page of a higher prestige pointing to i is more important than a page of a lower prestige pointing to i**
  - **In other words, a page is important if it is pointed to by other important pages**

PageRank algorithm

- According to **rank prestige**, the importance of page i (i's PageRank score) is
  - the sum of the PageRank scores of all pages that point to i

- Since a page may point to many other pages, its prestige score should be shared.
- The Web as a directed graph $G = (V, E)$
  - The PageRank score of the page i (denoted by $P(i)$) is defined by:

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j}$$

O_j is the number of out-link of j

Matrix notation

- Let $n = |V|$ be the total number of pages
- We have a system of n linear equations with n unknowns. We can use a matrix to represent them.
- Let P be a n-dimensional column vector of PageRank values, i.e., $P = (P(1), P(2), \ldots, P(n))^T$
- Let A be the adjacency matrix of our graph with

$$A_{ij} = \begin{cases} \frac{1}{O_i} & if\,(i,j) \in E \\ 0 & otherwise \end{cases}$$

We're considering all the elements of column j of the original matrix A

We have to transpose A in the matrix notation of PageRank

- We can write the n equations $P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j}$ with

$$P = A^T P \quad \textbf{(PageRank)}$$

Solve the PageRank equation

$$P = A^T P$$

- This is the characteristic equation of the **eigensystem**, where the solution to *P* is an **eigenvector** with the corresponding **eigenvalue** of 1
- It turns out that if ***some conditions*** are satisfied, 1 is the largest eigenvalue and the PageRank vector P is the **principal eigenvector**.
- A well known mathematical technique called **power iteration** can be used to find P
- **Problem:** the above Equation does not quite suffice because the Web graph does not meet the conditions.

Using Markov chain

- To introduce these conditions and the enhanced equation, let us derive the same above Equation based on the Markov chain.
  - In the Markov chain, each Web page or node in the Web graph is regarded as a state.
  - A hyperlink is a transition, which leads from one state to another state with a probability.
- This framework models Web surfing as a stochastic process.
- Random walk
  - It models a Web surfer randomly surfing the Web as state transition.

Random surfing

- **Recall we used O_i to denote the number of out-links of a node i**
- **Each transition probability is $1/O_i$ if we assume the Web surfer will click the hyperlinks in the page i uniformly at random.**
  - **the "back" button on the browser is not used**
  - **the surfer does not type in an URL**

Transition probability matrix

- **Let A be the state transition probability matrix:**

$$A = \begin{pmatrix} A_{11} & A_{12} & . & . & . & A_{1n} \\ A_{21} & A_{22} & . & . & . & A_{2n} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ A_{n1} & A_{n2} & . & . & . & A_{nn} \end{pmatrix}$$

- **A_{ij} represents the transition probability that the surfer in state i (page i) will move to state j (page j).**
- **Can A be the adjacency matrix previously discussed?**

Let us start

- **Given an initial probability distribution vector that a surfer is at each state (or page)**
  - $p_0 = (p_0(1), p_0(2), \ldots, p_0(n))^T$ **(a column vector)**
  - **an $n \times n$ transition probability matrix A**

  we have

$$\sum_{i=1}^{n} p_0(i) = 1$$

$$\sum_{j=1}^{n} A_{ij} = 1 \qquad \textbf{(1)}$$

- **If the matrix A satisfies Equation (1), we say that A is the stochastic matrix of a Markov chain**

Back to the Markov chain

- In a Markov chain, a question of common interest is:
  - What is the probability that, after m steps/transitions (with $m \rightarrow \infty$), a random process/walker reaches a state j independently of the initial state of the walk
- We determine the probability that the system (or the random surfer) is in state j after 1 step (1 transition) by using the following reasoning:

$$p_1(j) = \sum_{i=1}^{n} A_{ij}(1)\, p_0(i)$$

We're still considering all the elements of column j of matrix A
We have to transpose A in the matrix notation …

where $A_{ij}(1)$ is the probability of going from i to j after 1 step.

State transition

- We can write this in matricial form:

$$P_1 = A^T P_0$$

- In general, the probability distribution after k steps/transitions is:

$$P_k = A^T P_{k-1}$$

Stationary probability distribution

- **By the Ergodic Theorem of Markov chain**
  - a finite Markov chain defined by the stochastic matrix A has a unique stationary probability distribution if A is irreducible and aperiodic
- **The stationary probability distribution means that**
  - after a series of transitions p_k will converge to a steady-state probability vector π regardless of the choice of the initial probability vector p_0, i.e.,

$$\lim_{k \to \infty} P_k = \pi$$

PageRank again

- **When we reach the steady-state, we have $P_k = P_{k+1} = \pi$, and thus**

$$\pi = A^{\mathrm{T}} \pi$$

- **π is the principal eigenvector (the one with the maximum magnitude) of A^{T} with eigenvalue of 1**

- **In PageRank, π is used as the PageRank vector P:**

$$P = A^T P$$

Is $P = \pi$ justified?

- **Using the stationary probability distribution π as the PageRank vector is reasonable and quite intuitive because**
  - **it reflects the long-run probabilities that a random surfer will visit the pages.**
  - **a page has a high prestige if the probability of visiting it is high**

Back to the Web graph

- Now let us come back to the real Web context and see whether the above conditions are satisfied, i.e.,
  - whether *A* is a stochastic matrix and
  - whether it is irreducible and aperiodic.
- None of them is satisfied.
- Hence, we need to extend the ideal-case to produce the "actual PageRank" model.

A is a not stochastic matrix

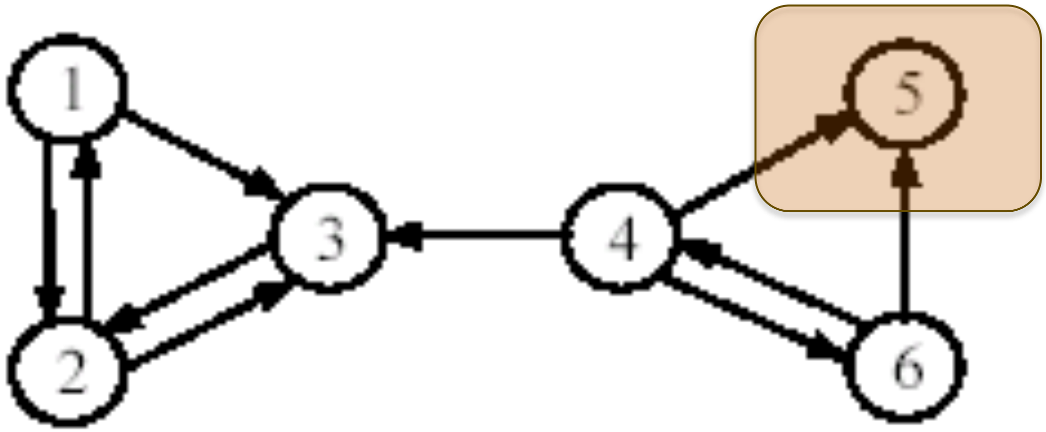
- A is the transition matrix of the Web graph

$$A_{ij} = \begin{cases} \frac{1}{O_i} & if\,(i,j) \in E \\ 0 & otherwise \end{cases}$$

- It does not satisfy equation: $\sum_{j=1}^{n} A_{ij} = 1$

  because many Web pages have no out-links, which are reflected in transition matrix *A* by some rows of complete 0's
  - Such pages are called the dangling pages (nodes).

An example Web hyperlink graph



$$A = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Fix the problem: two possible ways

1. **Remove** pages with **no out-links** during the PageRank computation
   - these pages do not affect the ranking of any other page directly
2. **Add** a complete set of **outgoing links** from each such page *i* to all the pages on the Web.

Let us use the 2nd method:

$$\overline{A} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

A is a not irreducible

- Irreducible means that the Web graph *G* is strongly connected

  Definition: A directed graph $G = (V, E)$ is strongly connected if and only if, for each pair of nodes $u, v \in V$, there is a directed path from u to v.

- A general Web graph represented by *A* is not irreducible because
  - for some pair of nodes u and v, there is no path from u to v
  - In our example, there is no directed path from nodes 3 to 4

A is a not aperiodic

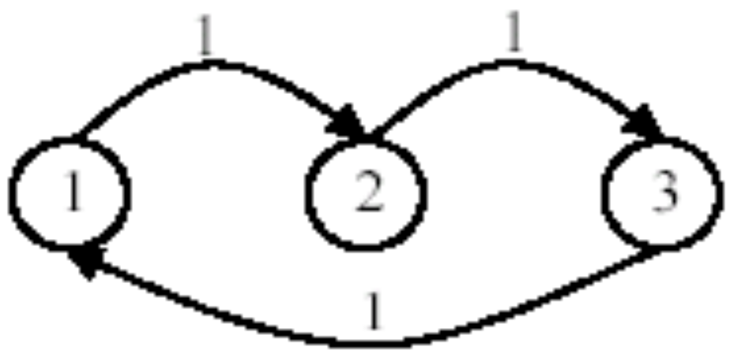
- A state i in a Markov chain being periodic means that there exists a directed cycle (from i to i) that a random walker traverses multiple times

- Definition: A state i is periodic with period $k > 1$ if k is the smallest number such that all paths leading from state i back to state i have a length that is a multiple of k
  - A Markov chain is aperiodic if all states are aperiodic.

An example: periodic

- This a periodic Markov chain with $k = 3$

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



- If we begin from state 1, to come back to state 1 the only path is 1-2-3-1 for some number of times, say h
  - Thus any return to state 1 will take $k \cdot h = 3h$ transitions.

Deal with irreducible and aperiodic matrices

- It is easy to deal with the above two problems with a single strategy.

- Add a link from each page to every page and give each link a small transition probability controlled by a parameter d (indeed, this probability will be $1-d$, where is also a probability)

- Obviously, the augmented transition matrix becomes irreducible and aperiodic
  - it becomes *irreducible* because it is strongly connected
  - it become *aperiodic* because we now have paths of all the possible lengths from state i back to state i

Improved PageRank

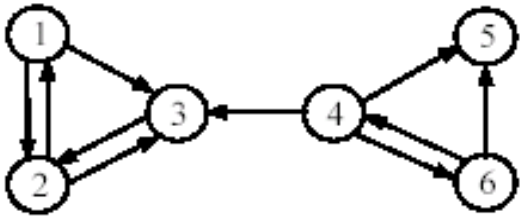
- After this augmentation, at a page, the random surfer has two options
  - With probability d, $0<d<1$, she randomly chooses an out-link to follow
  - With probability $1\text{-}d$, she stops clicking and jumps to a random page
- The following equation models the improved model:

$$\boldsymbol{P} = ((1-d)\frac{\boldsymbol{E}}{n} + dA^T)\boldsymbol{P}$$

n is important, since the matrix has to be *stochastic*

where $\boldsymbol{E}$ is a $n \times n$ square matrix of all 1's

Follow our example



The matrix made ***stochastic***, which is still:

- ***periodic*** (see state 3)
- ***reducible*** (no path from 3 to 4)

$$A = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Transposed matrix

$d = 0.9$

$$(1-d)\frac{E}{n} + dA^T = \begin{pmatrix} 1/60 & 7/15 & 1/60 & 1/60 & 1/6 & 1/60 \\ 7/15 & 1/60 & 11/12 & 1/60 & 1/6 & 1/60 \\ 7/15 & 7/15 & 1/60 & 19/60 & 1/6 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/6 & 7/15 \\ 1/60 & 1/60 & 1/60 & 19/60 & 1/6 & 7/15 \\ 1/60 & 1/60 & 1/60 & 19/60 & 1/6 & 1/60 \end{pmatrix}$$

The final PageRank algorithm

- $(1\text{-}d)E/n + dA^{\mathrm{T}}$ is a stochastic matrix (transposed). It is also irreducible and aperiodic
- Note that
  - $E = e\, e^{\mathrm{T}}$ where e is a column vector of 1's
  - $e^{\mathrm{T}} P = 1$ since P is the stationary probability vector π
- If we scale this equation:

$$P = ((1-d)\frac{E}{n} + dA^{T})P = (1-d)\frac{1}{n}e\, e^{T}P + dA^{T}P =$$

$$= (1-d)\frac{1}{n}e + dA^{T}P$$

by multiplying both sides by n, we have:

  - $e^{\mathrm{T}} P = n$ and thus:

$$P = (1-d)e + dA^{T}P$$

The final PageRank algorithm (cont …)

- **Given:**

$$\boldsymbol{P} = (1-d)\boldsymbol{e} + dA^T\boldsymbol{P}$$

$$A_{ji} = \begin{cases} \frac{1}{O_j} & if\,(j,i) \in E \\ 0 & otherwise \end{cases}$$

PageRank for each page *i* is:

$$P(i) = (1-d) + d\sum_{j=1}^{n} A_{ji}P(j)$$

that is equivalent to the formula given in the PageRank paper [BP98]

- **The parameter d is called the damping factor which can be set to between 0 and 1. $d = 0.85$ was used in the PageRank paper**

$$P(i) = (1-d) + d\sum_{(j,i)\in E} \frac{P(j)}{O_j}$$

[BP98] Sergey Brin and Lawrence Page. **The Anatomy of a Large-Scale Hypertextual Web Search Engine.** WWW Int.l Conf., 1998.

Compute PageRank

- Use the **power iteration** method

PageRank-Iterate(G)

$P_0 \leftarrow e/n$ — Initialization

$k = 0$

repeat

$P_{k+1} \leftarrow (1-d)\frac{e}{n} + dA^T P_k$;

$k = k + 1$;

until $\|P_{k+1} - P_k\|_1 < \varepsilon$ — Norm 1 less than 10^{-6}

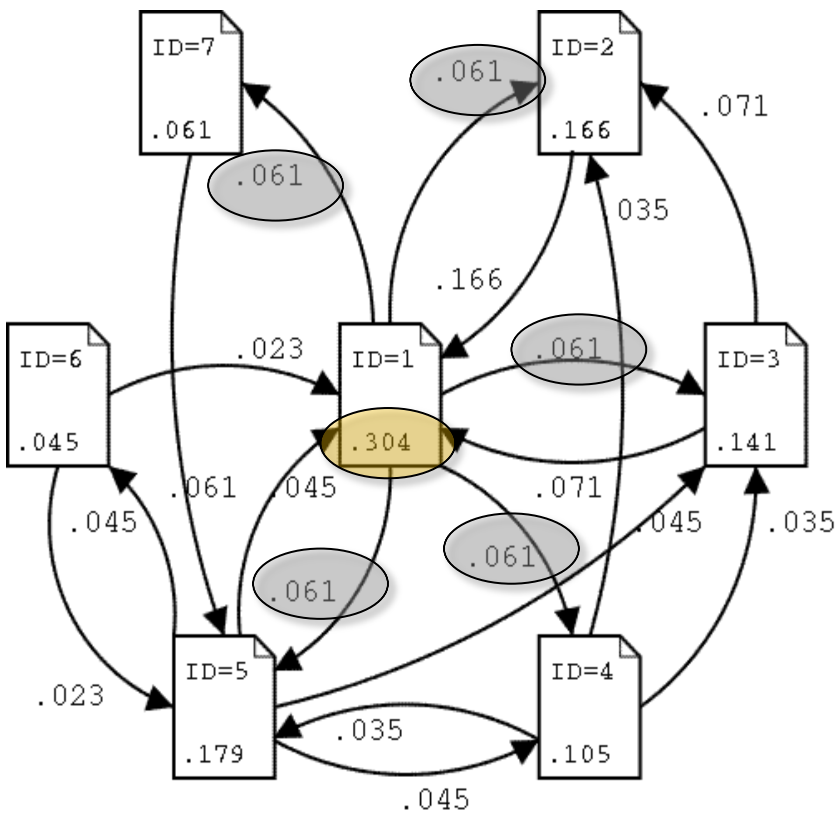
return P_{k+1}

Fig. 6. The power iteration method for PageRank

Again PageRank

- Without scaling the equation (by multiplying by n), we have $e^T P = 1$ (i.e., the sum of all PageRanks is one), and thus:

$$P(i) = \frac{1-d}{n} + d \sum_{(j,i)\in E} \frac{P(j)}{O_j}$$

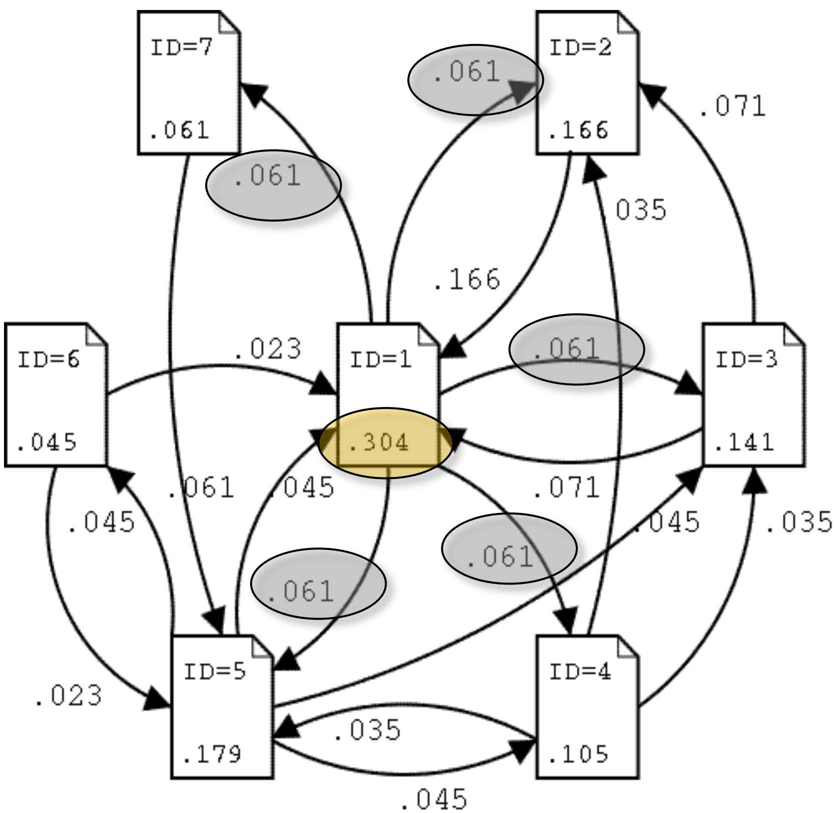


- Important pages
  - are cited/pointed by other important ones
- In the example, the most important is ID=1
  - P(ID=1) = 0.304
- P(ID=1) distributes is "rank" among all its 5 outgoing links
  - ID= 2, 3, 4, 5, 7
  - 0.304 = 0.061 * 5

Again PageRank

- Without scaling the equation (by multiplying by n), we have $e^T P = 1$ (i.e., the sum of all PageRanks is one), and thus:

$$P(i) = \frac{1-d}{n} + d \sum_{(j,i)\in E} \frac{P(j)}{O_j}$$



- The stationary probability **P(ID=1)** is obtained by:

  (1-d)/n +
  d (0.023+0.166+0.071+0.045)=

  (0.15)/7 +
  0.85(0.023+0.166+0.071+0.045)=

  0.304

Advantages of PageRank

- **Fighting spam**. A page is important if the pages pointing to it are important.
  - Since it is not easy for Web page owner to add in-links into his/her page from other important pages, it is thus not easy to influence PageRank.
- **PageRank is a global measure and is query independent**.
  - PageRank values of all the pages are computed and saved off-line rather than at query time.
- **Criticism**: Query-independence. It could not distinguish between pages that are authoritative in general and pages that are authoritative on the query topic.

Contents

- Introduction
- Network properties
- Social network analysis
- Co-citation and bibliographic coupling
- PageRank
- HITS
- Summary

HITS

- HITS stands for Hypertext Induced Topic Search
- Unlike PageRank which is a static ranking algorithm, HITS is query dependent

- When a user issues a search query, HITS
  - first expands the list of relevant pages returned by a search engine, and
  - then produces two rankings of the expanded set of pages, authority ranking and hub ranking
  - it can be exploited by a meta search engine, which re-ranks pages returned by one or many WSEs

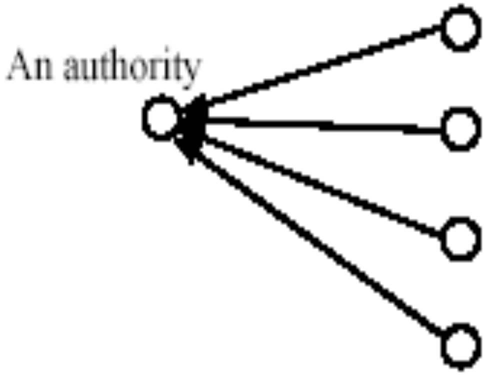
Authorities and Hubs

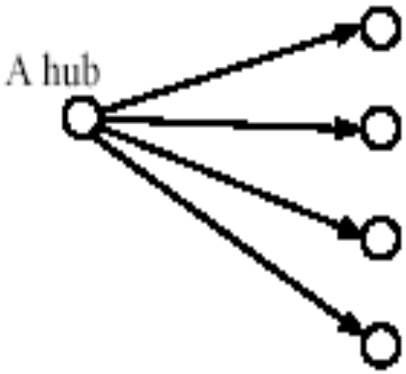
Authority: Roughly, an authority is a page with many in-links.

- **The idea is that the page may have good or authoritative content on some topic**
- **thus many people trust it and link to it.**

Hub: A hub is 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.**

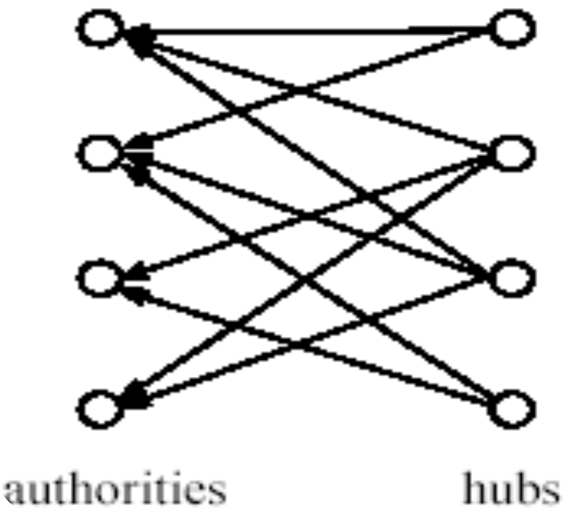




The key idea of HITS

- A good hub points to many good authorities
- A good authority is pointed to by many good hubs

- Authorities and hubs have a mutual reinforcement relationship !!
- Some densely linked authorities and hubs (a bipartite sub-graph):



The HITS algorithm: Grab pages

- **Given a broad search query, *q*, HITS collects a set of pages as follows:**
  - It sends query *q* to a search engine
  - It then collects the *t* (*t* = 200 is used in the HITS paper) highest ranked pages. This set is called the root set *W*
  - It then grows *W* by including any page pointed to by a page in *W* and any page that points to a page in *W*.
  - This gives a larger set *S*, base set

The link graph G

- **HITS works on the pages in S, and assigns every page in S an authority score and a hub score**

- **Given S, $|S|=n$, we again use *G* = (*V*, *E*) to denote the hyperlink graph of S**
- **We use L to denote the adjacency matrix of the graph.**

$$L_{ij} = \begin{cases} 1 & if\,(i,j) \in E \\ 0 & otherwise \end{cases}$$

The HITS algorithm

- ***a*(*i*)** : **authority score of page *i***
- ***h*(*i*)** : **hub score of page *i***

- **The mutual reinforcing relationship of the two scores is represented as follows:**

$$a(i) = \sum_{(j,i) \in E} h(j)$$

$$h(i) = \sum_{(i,j) \in E} a(j)$$

HITS in matrix form

- **Let $\boldsymbol{a}$ denote the column vector with all the authority scores**

  $$\boldsymbol{a} = (a(1), a(2), \ldots, a(n))^{\mathrm{T}}$$

- **Let $\boldsymbol{h}$ denote the column vector with all the authority scores**

  $$\boldsymbol{h} = (h(1), h(2), \ldots, h(n))^{\mathrm{T}}$$

- **Then:**

  $$\boldsymbol{a} = \boldsymbol{L}^{\mathrm{T}} \boldsymbol{h}$$

  $$\boldsymbol{h} = \boldsymbol{L}\, \boldsymbol{a}$$

Computation of HITS

- **The computation of authority scores and hub scores is the same as the computation of the PageRank scores, using power iteration.**
- **If we use a_k and h_k to denote authority and hub vectors at the k^{th} iteration, the iterations for generating the final solutions are:**

$$a_k = L^T L \, a_{k-1}$$

$$h_k = L \, L^T h_{k-1}$$

starting with:

$$a_0 = h_0 = (1,1,\ldots,1)$$

The algorithm

HITS-Iterate(G)

$\boldsymbol{a}_0 = \boldsymbol{h}_0 = (1, 1, \ldots, 1)$;
$k = 1$
Repeat

$$\boldsymbol{a}_k = \boldsymbol{L}^T \boldsymbol{L} \boldsymbol{a}_{k-1};$$

$$\boldsymbol{h}_k = \boldsymbol{L} \boldsymbol{L}^T \boldsymbol{h}_{k-1};$$

normalize $\boldsymbol{a}_k$;
normalize $\boldsymbol{h}_k$;
$k = k + 1$;

until $\boldsymbol{a}_k$ and $\boldsymbol{h}_k$ do not change significantly;
return $\boldsymbol{a}_k$ and $\boldsymbol{h}_k$

Fig. 9. The HITS algorithm based on power iteration

Relationships with co-citation and bibliographic coupling

- **Recall that co-citation of pages *i* and *j*, denoted by C_{ij}**
  - the authority matrix ($L^T L$) of HITS is the co-citation matrix C

$$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj} = (\mathbf{L}^T \mathbf{L})_{ij}$$

- **Recall the bibliographic coupling of two pages *i* and *j*, denoted by B_{ij}**
  - the hub matrix ($L L^T$) of HITS is the bibliographic coupling matrix B

$$B_{ij} = \sum_{k=1}^{n} L_{ik} L_{jk} = (\mathbf{L}\mathbf{L}^T)_{ij},$$

Strengths and weaknesses of HITS

- **Strength**: its ability to rank pages according to the query topic, which may be able to provide more relevant authority and hub pages.
- **Weaknesses**:
  – **It is easily spammed**. It is in fact quite easy to influence HITS since adding out-links in one's own page is so easy.
  – **Topic drift**. Many pages in the expanded set may be off-topic.
  – **Inefficiency at query time**: The query time evaluation is slow. Collecting the root set, expanding it and performing eigenvector computation are all expensive operations

Contents

Summary

- In this chapter, we introduced
  - Some important properties of the Web networks
  - Social network analysis, centrality and prestige
  - Co-citation and bibliographic coupling
  - PageRank, which powers Google
  - HITS
- Yahoo! and Bing have their own link-based algorithms as well, but not published.
- Important to note: Hyperlink based ranking is not the only algorithm used in search engines. In fact, it is combined with many content based factors to produce the final ranking presented to the user.

Summary

- **Links can also be used to find communities, which are groups of content-creators or people sharing some common interests.**
  - Web communities
  - Email communities
  - Named entity communities

- **Focused crawling: combining contents and links to crawl Web pages of a specific topic.**
  - Follow links and
  - Use learning/classification to determine whether a page is on topic.

Chakrabarti and Van den Berg. **"Focused crawling: a new approach to topic-specific Web resource discovery"**. Computer Networks, 1999 - Elsevier