

# Graph and Web Mining - Motivation, Applications and Algorithms



**PROF. EHUD GUIDES**  
**DEPARTMENT OF COMPUTER SCIENCE**  
**BEN-GURION UNIVERSITY, ISRAEL**

# Web mining - Outline

2

- **Introduction**
- **Web Content Mining**
- **Web usage mining**
- **Web Link and Structure Mining**
- **Web Crawler**



# Web Structure Mining

## Link Analysis Algorithms

3

**PAGE RANK**

# Introduction

4

- 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 “classification technique”, Google estimates: 10 million 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 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 ...)

5

- 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 ...)

6

- During 1997-1998, two most influential hyperlink based search algorithms **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 powers 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, 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..

# Social network analysis

8

- Social network is the study of social entities (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.
- From the network, we can study the properties of its structure, and **the role, position** and **prestige** of each social actor.
- We can also find various kinds of sub-graphs, e.g., **communities** formed by groups of actors.



# Social network and the Web

9

- Social network analysis is useful for the Web because the Web is essentially a virtual society, and thus a virtual social network,
  - Each page: a social actor and
  - each hyperlink: a relationship.
- Many results from social network can be adapted and extended for use 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

10

- **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

11

Central actors are the most active actors that have most links or ties with other actors. Let the total number of actors in the network be  $n$ .

**Undirected graph:** In an undirected graph, the **degree centrality** of an actor  $i$  (denoted by  $C_D(i)$ ) is simply the node degree (the number of edges) of the actor node, denoted by  $d(i)$ , normalized with the maximum degree,  $n-1$ .

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

**Directed graph:** In this case, we need to distinguish **in-links** of actor  $i$  (links pointing to  $i$ ), and **out-links** (links pointing out from  $i$ ). The degree centrality is defined based on only the out-degree (the number of out-links or edges),  $d_o(i)$ .

$$C'_D(i) = \frac{d_o(i)}{n-1} \quad (2)$$

# Closeness Centrality

12

This view of centrality is based on the closeness or distance. The basic idea is that an actor  $x_i$  is central if it can easily interact with all other actors. That is, its distance to all other actors is short. Thus, we can use the shortest distance to compute this measure. Let the shortest distance from actor  $i$  to actor  $j$  be  $d(i, j)$ .

**Undirected graph:** The closeness centrality  $C_C(i)$  of actor  $i$  is defined as

$$C_C(i) = \frac{n-1}{\sum_{j=1}^n d(i, j)} \quad (3)$$

The value of this measure also ranges between 0 and 1 as  $n-1$  is the minimum value of the denominator, which is the sum of shortest distances from  $i$  to all other actors. Note that this equation is only meaningful for a connected graph.

**Directed graph:** The same equation can be used for a directed graph. The distance computation needs to consider directions of links or edges.

# Betweenness Centrality

13

- If two non-adjacent actors  $j$  and  $k$  want to interact and 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.

# Betweenness Centrality (cont ...)

14

- **Undirected graph:** Let  $p_{jk}$  be the number of shortest paths between actor  $j$  and actor  $k$ .
- The betweenness of an actor  $i$  is defined as the number of shortest paths that pass  $i$  ( $p_{jk}(i)$ ) normalized by the total number of shortest paths.

$$\sum_{j < k} \frac{p_{jk}(i)}{p_{jk}} \quad (4)$$

# Betweenness Centrality (cont ...)

15

Note that there may be multiple shortest paths between  $j$  and  $k$ . Some passes  $i$  and some do not. If we are to ensure the value range is between 0 and 1, we can normalize it with  $(n-1)(n-2)/2$ , which is the maximum value of the above quantity, i.e., the number of pairs of actors not including  $i$ . The final betweenness of  $i$  is defined as

$$C_B(i) = \frac{2 \sum_{j < k} \frac{p_{jk}(i)}{P_{jk}}}{(n-1)(n-2)} \quad (5)$$

Unlike the closeness measure, the betweenness can be computed even if the graph is not connected.

**Directed graph:** The same equation can be used but must be multiplied by 2 because there are now  $(n-1)(n-2)$  pairs considering a path from  $j$  to  $k$  is different from a path from  $k$  to  $j$ . Likewise,  $p_{jk}$  must consider paths from both directions.

# Prestige

16

- Prestige is a more refined measure of prominence of an actor than centrality.
  - Distinguish: 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 use only in-links.
- **Difference between centrality and prestige:**
  - centrality focuses on out-links
  - 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

17

Based on the definition of the prestige, it is clear that an actor is prestigious if it receives many in-links or nominations. Thus, the simplest measure of prestige of an actor  $i$  (denoted by  $P_D(i)$ ) is its in-degree.

$$P_D(i) = \frac{d_I(i)}{n-1}, \quad (6)$$

where  $d_I(i)$  is in-degree of  $i$  (the number of in-links of actor  $i$ ) and  $n$  is the total number of actors in the network. As in the degree centrality, dividing  $n - 1$  standardizes the prestige value to the range from 0 and 1. The maximum prestige value is 1 when every other actor links to or chooses actor  $i$ .

# Proximity prestige

18

- The degree index of prestige of an actor  $i$  only considers the actors that are adjacent to  $i$ .
- The **proximity prestige** generalizes it by considering both the actors directly and indirectly linked to actor  $i$ .
  - 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 as closeness or distance of other actors to  $i$ .
- Let  $d(j, i)$  denote the distance from actor  $j$  to actor  $i$ .

# Proximity prestige (cont ...)

19

$$\frac{\sum_{j \in I_i} d(j, i)}{|I_i|}, \quad (7)$$

where  $|I_i|$  is the size of the set  $I_i$ . If we look at the ratio or proportion of actors who can reach  $i$  to the average distance that these actors are from  $i$ , we obtain the following, which has the value range of  $[0, 1]$ :

$$P_p(i) = \frac{|I_i|/(n-1)}{\sum_{j \in I_i} d(j, i) / |I_i|}, \quad (8)$$

where  $|I_i|/(n-1)$  is the proportion of actors that can reach actor  $i$ . In one extreme, every actor can reach actor  $i$ , which gives  $|I_i|/(n-1) = 1$ . The denominator is 1 if every actor is adjacent to  $i$ . Thus,  $P_p(i) = 1$ . On the other extreme, no actor can reach actor  $i$ . Then  $|I_i| = 0$ , and  $P_p(i) = 0$ . Each link has the unit distance.

# Rank prestige

20

- In the previous two prestige measures, an important factor is 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 chosen by a less important person.
  - For example, if a company CEO votes for a person is much more important than a worker votes for the 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 defined as a linear combination of links that point to  $i$ :

$$P_R(i) = A_{1i}P_R(1) + A_{2i}P_R(2) + \dots + A_{ni}P_R(n), \quad (9)$$

where  $A_{ji} = 1$  if  $j$  points to  $i$ , and 0 otherwise. This equation says that an actor's rank prestige is a function of the ranks of the actors who vote or choose the actor, which makes perfect sense.

Since we have  $n$  equations for  $n$  actors, mathematically we can write them in the matrix notation. We use  $\mathbf{P}$  to represent the vector that contains all the rank prestige values, i.e.,  $\mathbf{P} = (P_R(1), P_R(2), \dots, P_R(n))^T$  ( $T$  means **matrix transpose**).  $\mathbf{P}$  is represented as a column vector. We use matrix  $\mathbf{A}$  (where  $A_{ij} = 1$  if  $i$  points to  $j$ , and 0 otherwise) to represent the adjacency matrix of the network or graph. As a notational convention, we use bold italic letters to represent matrices. We then have

$$\mathbf{P} = \mathbf{A}^T \mathbf{P} \quad (10)$$

This equation is precisely the characteristic equation used for finding the **eigensystem** of the matrix  $\mathbf{A}^T$ .  $\mathbf{P}$  is an **eigenvector** of  $\mathbf{A}^T$ .

# PageRank

22

**LARRY PAGE AND SERGEY BRIN WWW7**

**CITED BY 6561 PAPERS!**

# Simple recursive formulation

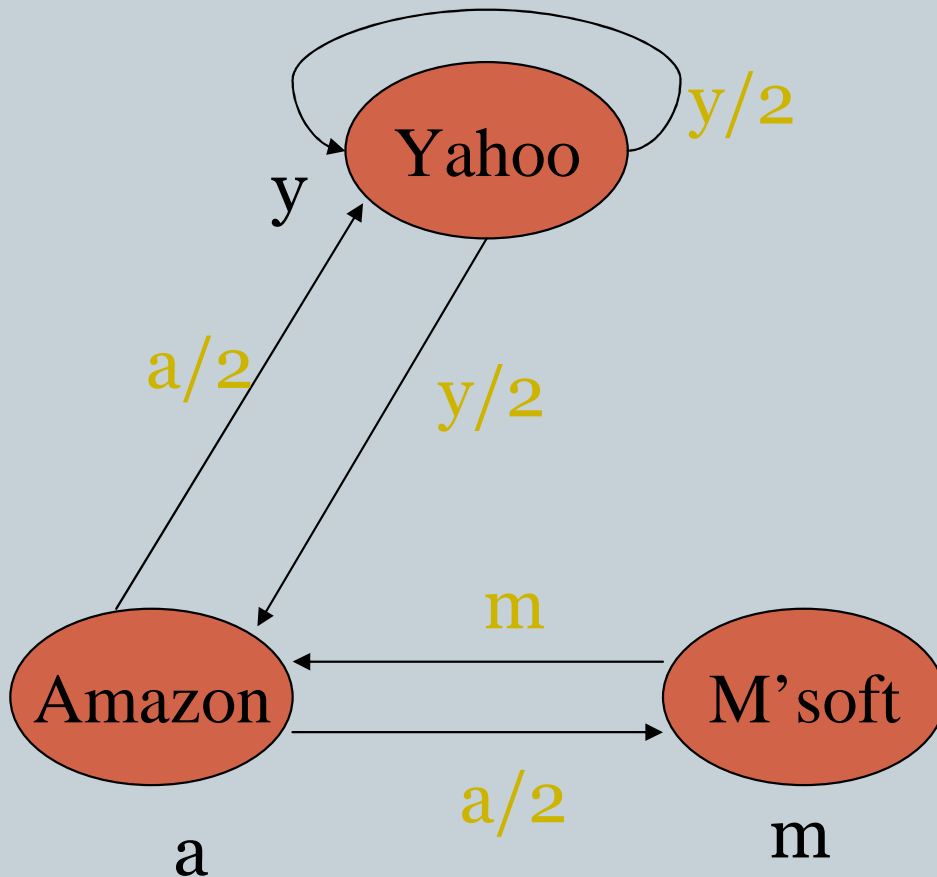
23

- Each link's vote is proportional to the **importance** of its source page
- If page **P** with importance **x** has **n** outlinks, each link gets  **$x/n$**  votes
- Page **P**'s own importance is the sum of the votes on its inlinks.

# Simple “flow” model

24

The web in 1983



$$y = y/2 + a/2$$

$$a = y/2 + m$$

$$m = a/2$$



# Solving the flow equations

25

- 3 equations, 3 unknowns, no constants
  - No unique solution
  - All solutions equivalent modulo scale factor
- Additional constraint forces uniqueness
  - $y+a+m = 1$
  - $y = 2/5, a = 2/5, m = 1/5$
- Gaussian elimination method works for small examples, but we need a better method for large graphs

# Matrix formulation

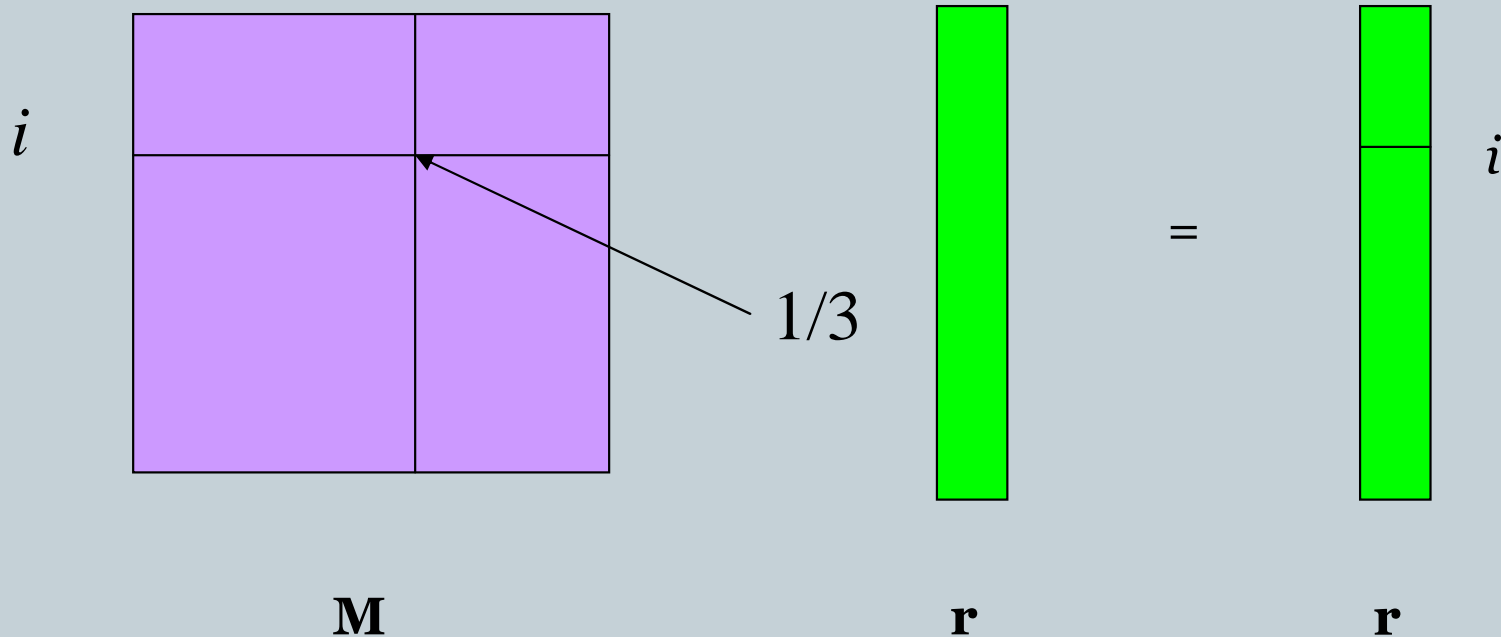
26

- Matrix  $\mathbf{M}$  has one row and one column for each web page
- Suppose page  $j$  has  $n$  outlinks
  - If  $j \neq i$ , then  $M_{ij} = 1/n$
  - Else  $M_{ij} = 0$
- $\mathbf{M}$  is a **column stochastic matrix**
  - A **column stochastic matrix** is a square matrix whose columns consist of nonnegative real numbers whose sum is 1.
- Suppose  $\mathbf{r}$  is a vector with one entry per web page
  - $r_i$  is the importance score of page  $i$
  - Call it the **rank vector**
  - $|\mathbf{r}| = 1$

# Example

27

Suppose page  $j$  links to 3 pages, including  $i$

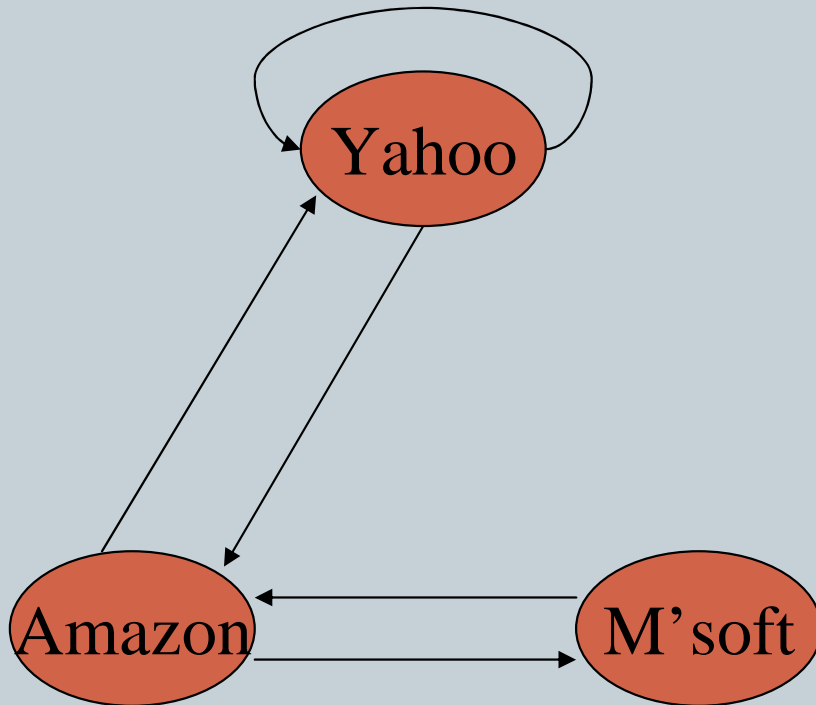


The flow equations can be written

$$r = Mr$$

# Example

28



$$y = y/2 + a/2$$

$$a = y/2 + m$$

$$m = a/2$$

	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

$$\mathbf{r} = \mathbf{M}\mathbf{r}$$

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

# Random Walk Interpretation

29

- A **random walk** is a mathematical formalization of a trajectory that consists of taking successive random steps.
- Imagine a **random web surfer**
  - At any time  $t$ , surfer is on some page  $P$
  - At time  $t+1$ , the surfer follows an outlink from  $P$  uniformly at random
  - Ends up on some page  $Q$  linked from  $P$
  - Process repeats indefinitely
- Let  $\mathbf{p}(t)$  be a vector whose  $i^{\text{th}}$  component is the probability that the surfer is at page  $i$  at time  $t$ 
  - $\mathbf{p}(t)$  is a probability distribution on pages

# The stationary distribution

30

- Where is the surfer at time  $t+1$ ?
  - Follows a link uniformly at random
  - $\mathbf{p}(t+1) = \mathbf{M}\mathbf{p}(t)$
- Suppose the random walk reaches a state such that  $\mathbf{p}(t+1) = \mathbf{M}\mathbf{p}(t) = \mathbf{p}(t)$ 
  - Then  $\mathbf{p}(t)$  is called a **stationary distribution** for the random walk
- Our rank vector  $\mathbf{r}$  satisfies  $\mathbf{r} = \mathbf{M}\mathbf{r}$ 
  - So it is a stationary distribution for the random surfer

# Existence and Uniqueness

31

A central result from the theory of random walks (aka Markov processes):

For graphs that satisfy certain conditions, the stationary distribution is unique and eventually will be reached no matter what the initial probability distribution at time  $t = 0$ .

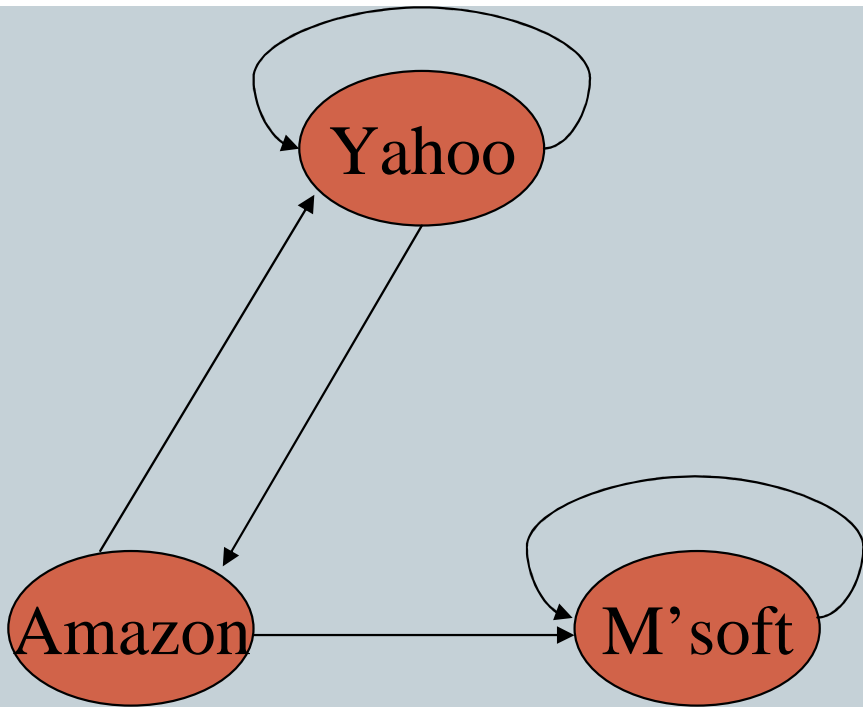
# Spider traps

32

- A group of pages is a **spider trap** if there are no links from within the group to outside the group
  - Random surfer gets trapped
- Spider traps violate the conditions needed for the random walk theorem



# Microsoft becomes a spider trap



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	1

	y	1	1	3/4	5/8		0
a	=	1	1/2	1/2	3/8	...	0
m		1	3/2	7/4	2		3

# Random teleports

34

- The Google solution for spider traps:
- At each time step, the random surfer has two options:
  - With probability  $\beta$ , follow a link at random
  - With probability  $1-\beta$ , jump to some page uniformly at random
  - Common values for  $\beta$  are in the range 0.8 to 0.9
- Surfer will teleport out of spider trap within a few time steps

# Matrix formulation

35

- Suppose there are  $N$  pages
  - Consider a page  $j$ , with set of outlinks  $O(j)$
  - We have  $M_{ij} = 1/|O(j)|$  when  $j \neq i$  and  $M_{ij} = 0$  otherwise
  - The random teleport is equivalent to
    - ✦ adding a **teleport link** from  $j$  to every other page with probability  $(1-\beta)/N$
    - ✦ reducing the probability of following each outlink from  $1/|O(j)|$  to  $\beta/|O(j)|$

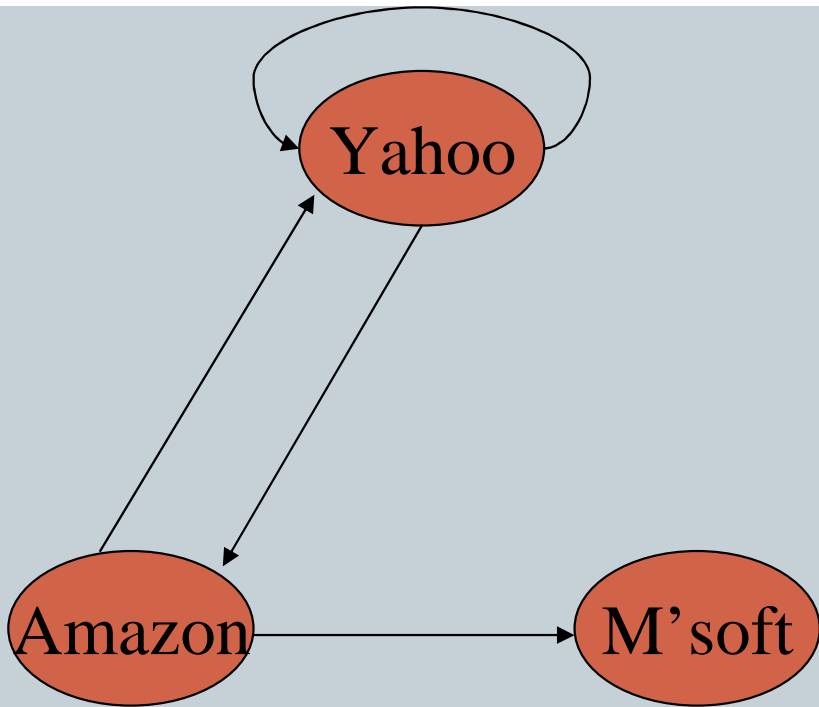
# Dead ends

36

- Pages with no outlinks are “dead ends” for the random surfer
  - Nowhere to go on next step

# Microsoft becomes a dead end

37



$$0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix} + 0.2 \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\begin{matrix} y \\ a \\ m \end{matrix} \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 1/15 \end{bmatrix}$$



Non-stochastic!

y	=	1	1	0.787	0.648	0
a		1	0.6	0.547	0.430	...
m		1	0.6	0.387	0.333	0

# Dealing with dead-ends

38

- **Teleport**
  - Follow random teleport links with probability 1.0 from dead-ends
  - Adjust matrix accordingly
- **Prune and propagate**
  - Preprocess the graph to eliminate dead-ends
  - Might require multiple passes
  - Compute page rank on reduced graph
  - Approximate values for deadends by propagating values from reduced graph

# HITS - Hypertext Induced Topic Selection

39

- Authorities - pages that contain useful information about the query topic
- Hubs - contain pointers to good information sources.
- Associating each page  $x$  with a hub score  $H(x)$  and an authority score  $A(x)$

$$H_{i+1}(x) = \sum_{(x,s)} A_i(s)$$

$$A_{i+1}(x) = \sum_{(p,x)} H_i(p)$$

- $(x,y)$  - hyperlink from  $x$  to  $y$
- $A_0(x) = H_0(x) = 1.0$
- Each iteration results are normalized

# HITS - Hypertext Induced Topic Selection

40

- Kleinberg (1999) was able to prove that this algorithm will always converge, and practical experience shows that it will typically do so within a few iterations
- HITS has been used for identifying relevant documents for topics in web catalogues and for implementing a “Related Pages” functionality
- The main drawback of the HITS algorithm is that the hubs and authority score must be computed iteratively from the query result while search query results need to be real time



# Structural Analysis of the Web

41

**PRATYUS PATNAIK, SUDIP SANYAL**  
**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,**  
**ALLAHABAD, INDIA**

# Main goal

42

- *to show that the Web is a Fractal.*
- *Each structurally isomorphic subgraph shows the same characteristics as the Web and follows the classical **Bow-tie model**.*

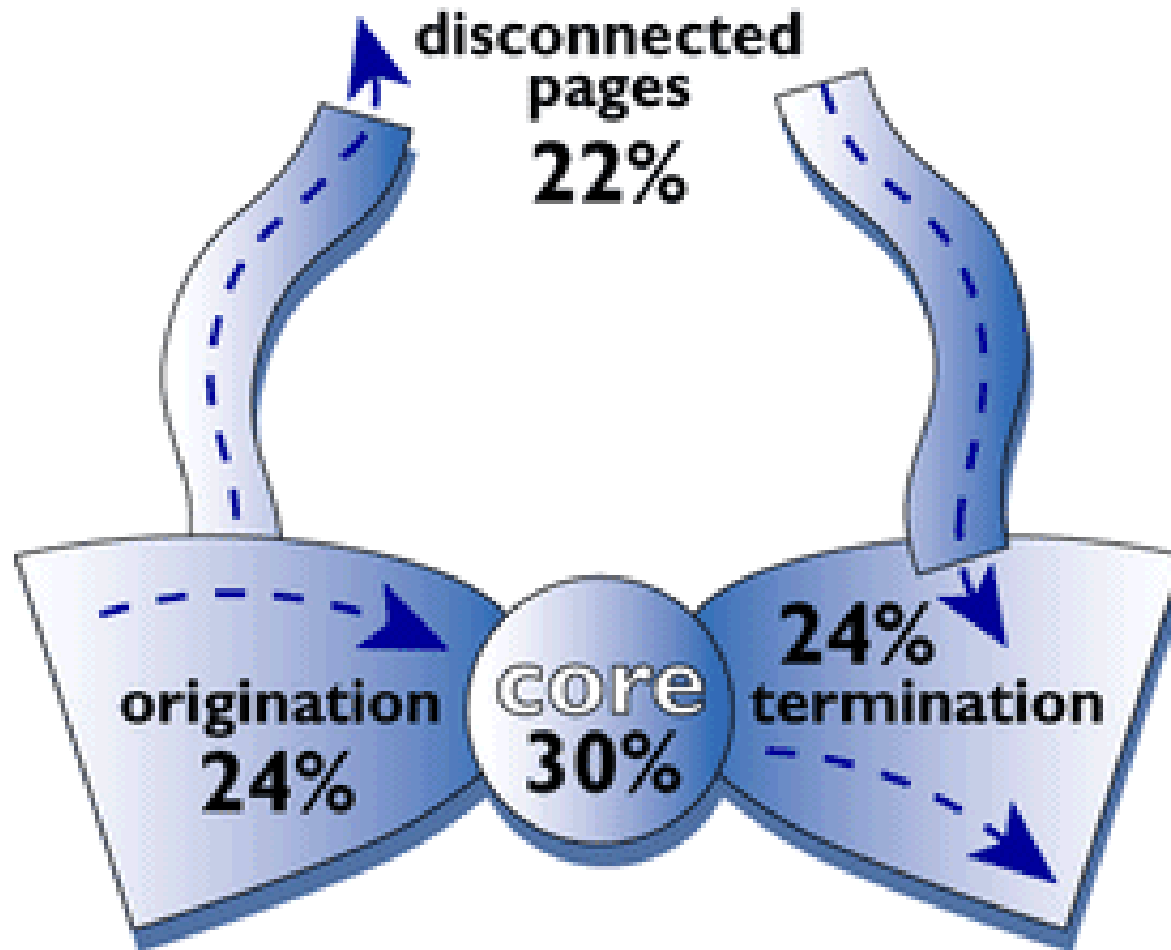
# Bow Tie Model

43

- the most common Bow Tie model consists of three main groups of web pages.
- **strongly connected** pages such that there is a path from any page within the core to any other page.
- **the "Origination"** consists of all pages that link to the strongly-connected core, but which have no links from the core back out to them.
- **the "Termination"** is the group of all pages that the strongly-connected core links to, but which have no links back into the core.

# Bow Tie Model

44



# The use of sub graph isomorphism

45

- “ we believe to capture the true insights on the structure of the web, we need to make use of pure graph based sub graph isomorphism algorithms. We applied iterative subgraph isomorphism algorithm on the webgraph to get the subgraphs. We then calculated various graph analysis parameter for those subgraphs. We found that each structurally similar subregion shows the same characteristic as the web and this holds for a number of parameters.”

# Terminologies and Algorithm

46

- **Web graph** – page = nodes , links =edges.
- **Graph Analysis Parameters** –
  - **Characteristic Path Length and Diameter** - the median of the means of the shortest paths from each node to every other node.
  - **Clustering Coefficient** - It is defined as the mean of the clustering indices of all the nodes in the graph. To find it, we find the neighbors of the node and then find the number of existing links amongst them. The ratio of the number of existing links to the number of possible links gives the clustering index of the node.

# Terminologies and Algorithm

47

- **Centrality and Centralization** - The degree centrality for a node is defined as:

$$C'_D(p_k) = \frac{\sum_{i=1}^n (a(p_i, p_k))}{n - 1}$$

- where  $a(p_i, p_k)$  is 1 iff  $p_i$  and  $p_k$  are directly connected in the direction from  $p_i$  to  $p_k$ . The degree centrality of a point is useful as an index of a potential communication ability.

# Terminologies and Algorithm

48

- **Degree Centralization** - The centralization of a network is calculated as the ratio of the centrality of each node of the network with a star network of the same size.
- **Betweenness Centrality** - It is based upon the frequency with which a point falls between pairs of other points on the shortest or geodesic paths connecting them.
- **Closeness Centrality**- It is related to the control of communication in a somewhat different manner. A point is viewed as central to the extent that it can avoid the control potential of others.



# Web Graph Characteristics

49

- Small World Network and Scale Invariance are two important characteristics reported in earlier works
- **Small World Network** : It is a complex network in which the distribution of connectivity is not confined to a certain scale, and where every node can be reached from every other by a small number of hops or steps.
- **Scale-free networks** usually contain centrally located and interconnected high degree nodes, which influence the way the network operates. For example, random node failures have very little effect on a scale-free network's connectivity or effectiveness; but deliberate attacks on such a node can lead to a complete break down.

## Conclusion - entirely structural point of view

50

- The Web is a fractal - It has cohesive sub-regions, at various scales, which exhibit the similar characteristics as the web for a lot of parameters.
- Each isomorphic subgraph nearly follows the classical Bow-Tie structure, with a robust core. This scalefree structural self similarity in the Web holds the key to building the theoretical models for understanding the evolution of the World Wide Web.

# Web Crawler

51

# Many names

52

- Crawler
- Spider
- Robot (or bot)
- Web agent
- Wanderer, worm, ...
- And famous instances: googlebot, scooter, slurp, msnbot, ...

# Web Crawler: Introduction

53

- Web Crawler (spider, robot) is a program which fetches information from the World Wide Web in a automated manner
  - It is mainly used to create a copy of all the visited pages for later processing (indexing and retrieving) by a search engine
  - It is also used to gather specific types of information from WWW
    - harvest email address (for spam purpose)
    - Event extraction
      - infectious disease outbreaks detection
- **In summary, Web Crawler is to finding, checking, and gathering stuff.**

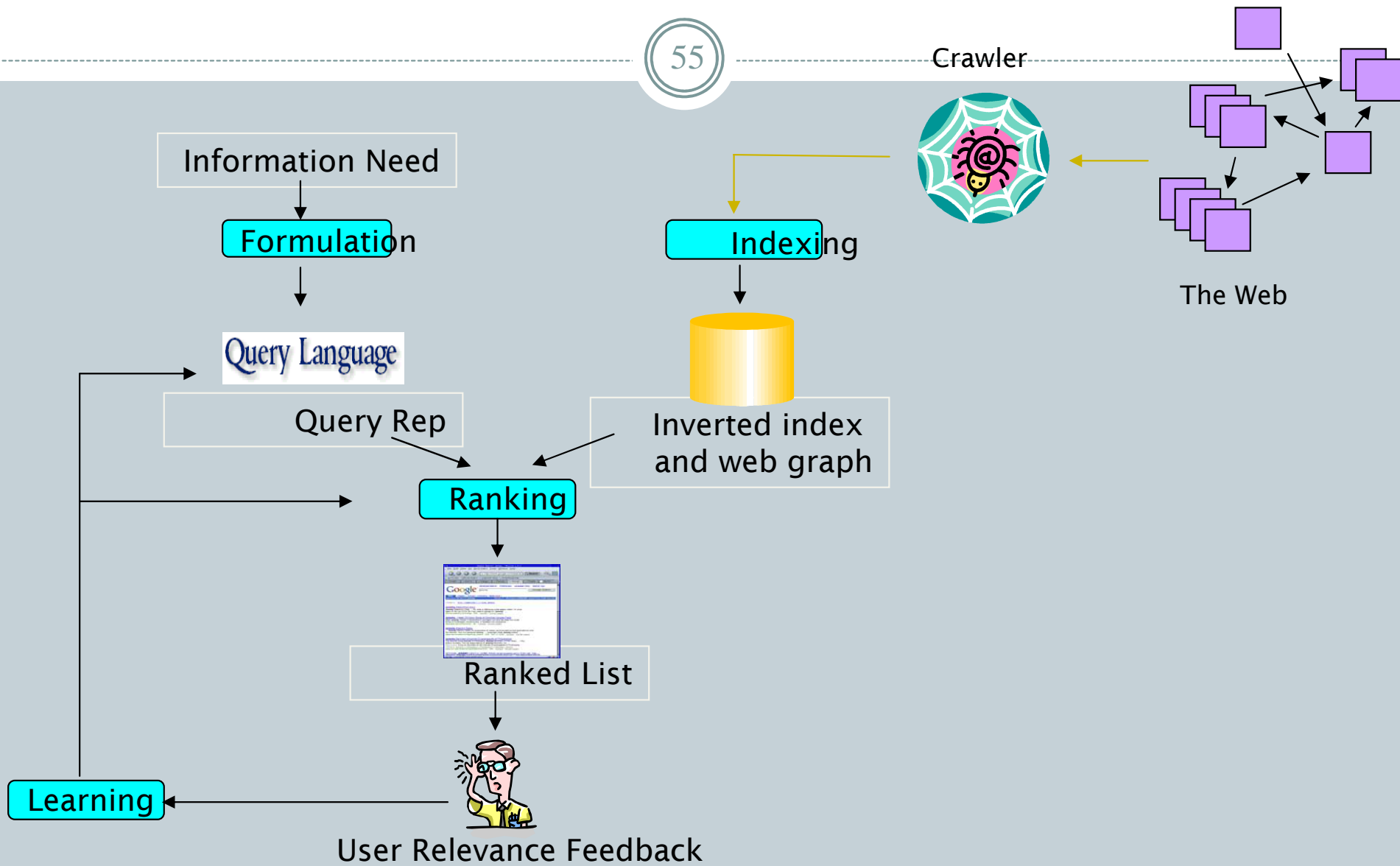
# Motivation for crawlers

54

- Support universal search engines (Google, Yahoo, MSN/Windows Live, Ask, etc.)
- Vertical (specialized) search engines, e.g. news, shopping, papers, recipes, reviews, etc.
- Business intelligence: keep track of potential competitors, partners
- Monitor Web sites of interest
- Evil: harvest emails for spamming, phishing...
- ... Can you think of some others?...

# Web Search Process

55



# Basic Crawling Algorithm

56

- G: a group of seed URLs
- Repeat:
  - choose a URL  $U$  from  $G$  //Crawling strategy
  - download the webpage  $w$  from  $U$
  - for each link  $l$  embedded in  $w$ 
    - if  $l$  has not been crawled before,
    - add  $l$  to  $G$ .
  -



# Robots Exclusion

57

- **The Robots Exclusion Protocol**

A Web site administrator can indicate which parts of the site should not be visited by a robot, by providing a specially formatted file on their site, in `http://.../robots.txt`.

- **The Robots META tag**

A Web author can indicate if a page may or may not be indexed, or analyzed for links, through the use of a special HTML META tag.

# The Robots META tag

58

- The Robots META tag allows HTML authors to indicate to visiting robots if a document may be indexed, or used to harvest more links. No server administrator action is required.
- For example: `<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">` a robot should neither index this document, nor analyze it for links.
- Currently only a few robots implement this.

# Robot Traps

59

- Because there is no editorial control over the internet, Web Crawlers should protect themselves from ill-formed html or misleading sites.
  - ill-formed html: page with 68 kB of null characters
  - misleading sites: CGI scripts can be used to generate infinite number of pages dynamically.
- Solutions
  - Eliminate URLs with non-textual data types
  - URL length check
  - maintain the statistics of a website. If the pages from a website exceedingly large, then remove the URLs coming from this website.

# Focused Crawler

60

- Generally speaking, focused crawler only crawls a restricted target space of Web pages
  - that may be of some “type” (e.g., homepages)
  - that may be of some “topic” (e.g., web mining)
- More specifically, focused crawler should be able to determine
  - How to decide whether a downloaded page is on-topic, or not?
  - How to choose the next URL to visit?

## Focused Crawler: Determine the next URL to VISIT

61

- **Hard-focus crawling:**
  - If a downloaded page is off-topic, stops following hyperlinks from this page.
- **Soft-focus crawling:**
  - obtains a page's relevance score (a score on the page's relevance to the target topic)
  - assigns this score to every URL extracted from this particular page, and adds to the priority queue


# Important References



- [1] Bing Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer publishing, 2009
  
- [2] Sergey Brin and Larry page, The anatomy of a web engine, WWW7, 1998,  
( better see the AMS version )


# Course Outline

63

- Searching Graphs and Related algorithms
  - Sub-graph isomorphism (Sub-sea)
  - Indexing and Searching – graph indexing
  - A new sequence mining algorithm
- Web mining and other applications
  - Document classification
  - Web mining
  - Short student presentation on their projects/papers 
- Conclusions

# Course Outline

64

- Searching Graphs and Related algorithms
  - Sub-graph isomorphism (Sub-sea)
  - Indexing and Searching – graph indexing
  - A new sequence mining algorithm
- Web mining and other applications
  - Document classification
  - Web mining
  - Short student presentation on their projects/papers
- **Conclusions** 



# Conclusions

65

- Graph mining is an interesting research area with many important applications
- The algorithms for graph mining are not trivial and require some effort for understanding them, but they often contain beautiful ideas
- Graph searching is a very current and hot research area which uses graph mining
- Web mining is a huge area with many applications. Web structure and link analysis often use graph mining algorithms

Thank You!