



HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.



ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

WEB MINING

LECTURE 05: LINK ANALYSIS (1/2)

ONE LOVE. ONE FUTURE.

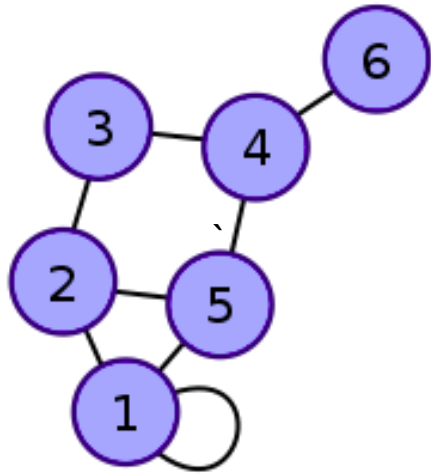
Main Problems in Link Analysis

- Graph Ranking: Analyze the role of nodes in graph
- Community detection: Detect communities consisting of members of similar nature
- Link prediction: Predicting the evolution of a graph over time
- Graph classification: Classify the vertices and edges of the graph into given classes

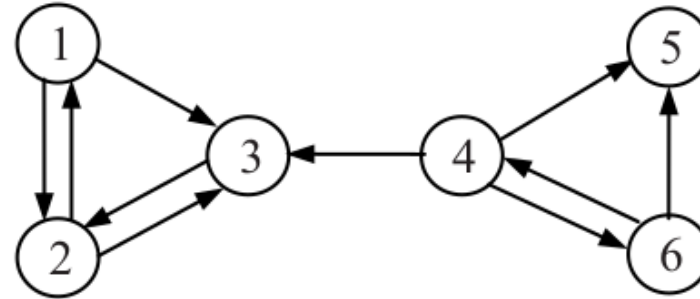
Agenda

1. Graph Ranking
2. Community Detection
3. Graph Representation

1. Graph Ranking/ 1.1 Basic concepts of graphs



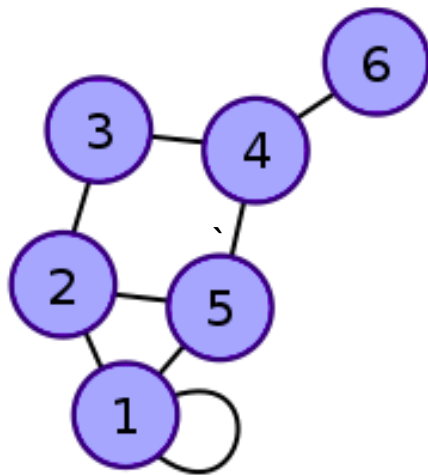
a) Undirected graph



Directed graph

Adjacency Matrix

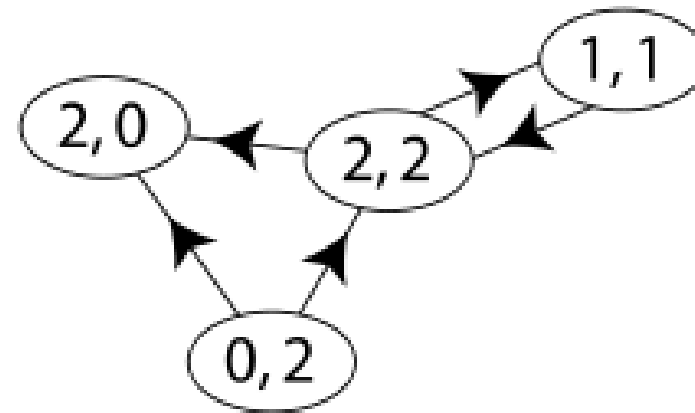
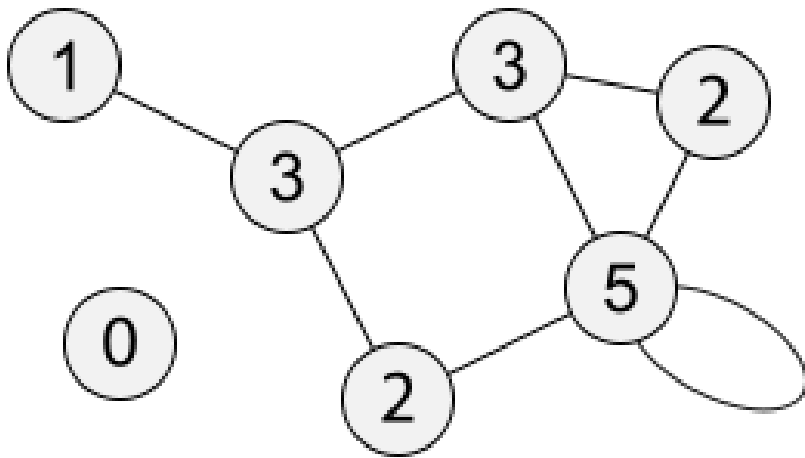
$$a[i, j] = \begin{cases} 1 & \text{if there is edge (i,j)} \\ 2 & \text{if there is a edge from a node to itself} \\ 0 & \text{otherwise} \end{cases}$$



$$\begin{pmatrix} 2 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Degree of a node

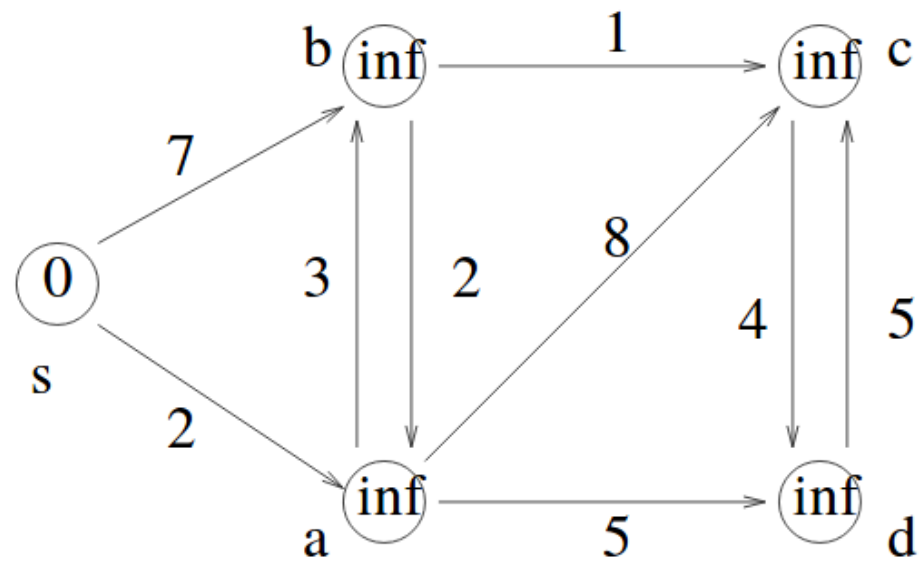
- $d_i(i)$ = number of in-edge of node i
- $d_o(i)$ = number of out-edge of node i



1.2 Dijkstra algorithm

- Find shortest path from source node s to the other nodes of graph
 - $d(v)$: Distance from node s to node v
 - S1**: Initialize $d(s) = 0$; $d(v) = \infty$
 - S2**: Arrange the nodes in a specific order in a queue Q
 - S3**: Get node u from queue Q then update distance $d(v)$ (if needed) of every node v adjacent to node u
- Go back to step **S2** until every node is computed

Example

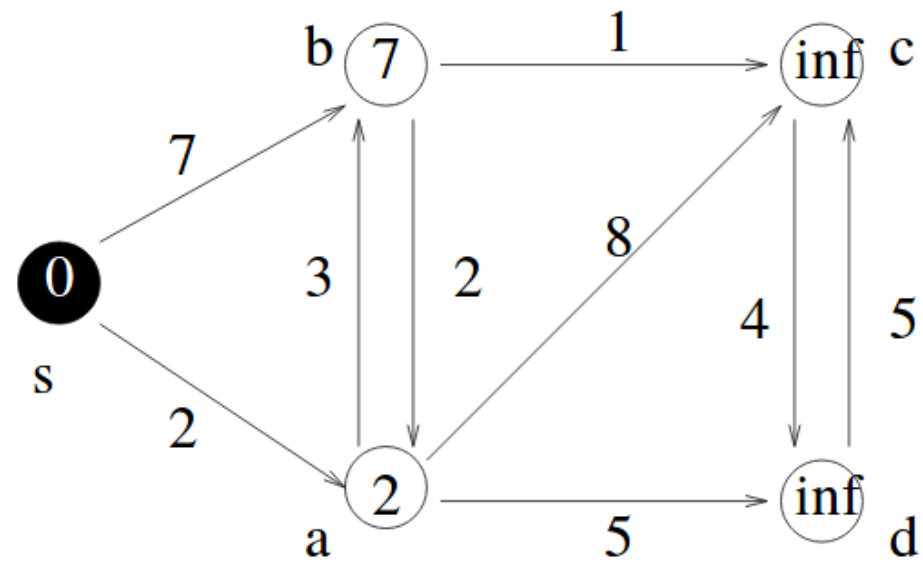


Example (cont.)

v	s	a	b	c	d
$d[v]$	0	∞	∞	∞	∞
$pred[v]$	nil	nil	nil	nil	nil
$color[v]$	W	W	W	W	W

v	s	a	b	c	d
$d[v]$	0	∞	∞	∞	∞

Example (cont.)

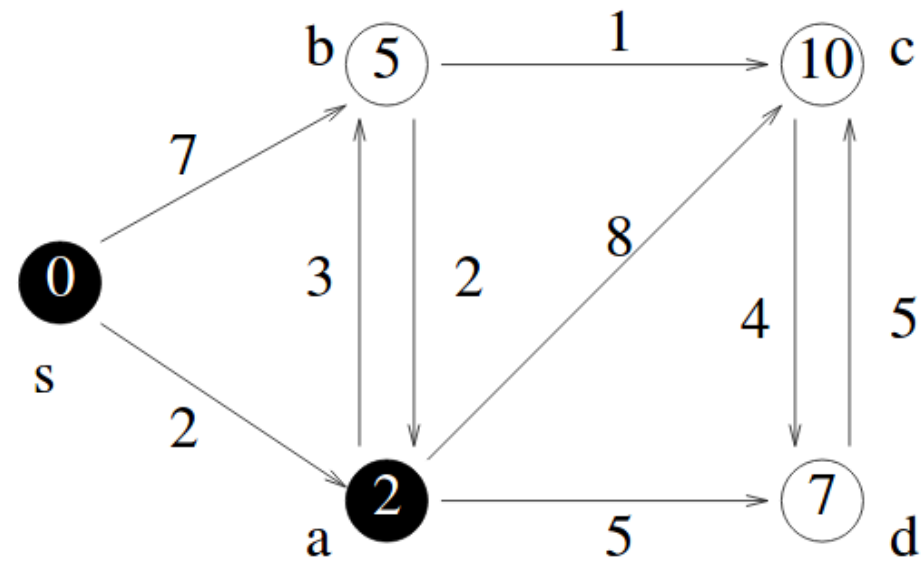


Example (cont.)

v	s	a	b	c	d
$d[v]$	0	2	7	∞	∞
$pred[v]$	nil	s	s	nil	nil
$color[v]$	B	W	W	W	W

v	a	b	c	d
$d[v]$	2	7	∞	∞

Example (cont.)

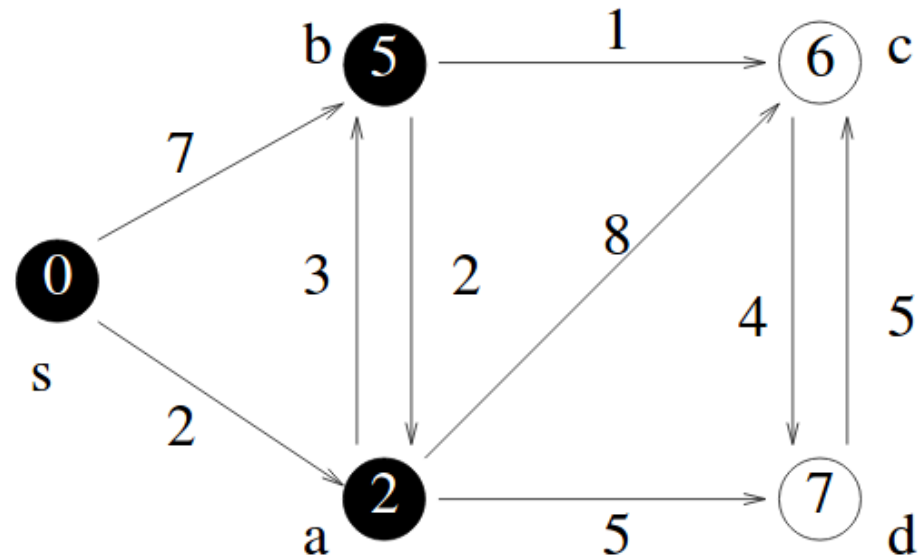


Example (cont.)

v	s	a	b	c	d
$d[v]$	0	2	5	10	7
$pred[v]$	nil	s	a	a	a
$color[v]$	B	B	W	W	W

v	b	c	d
$d[v]$	5	10	7

Example (cont.)

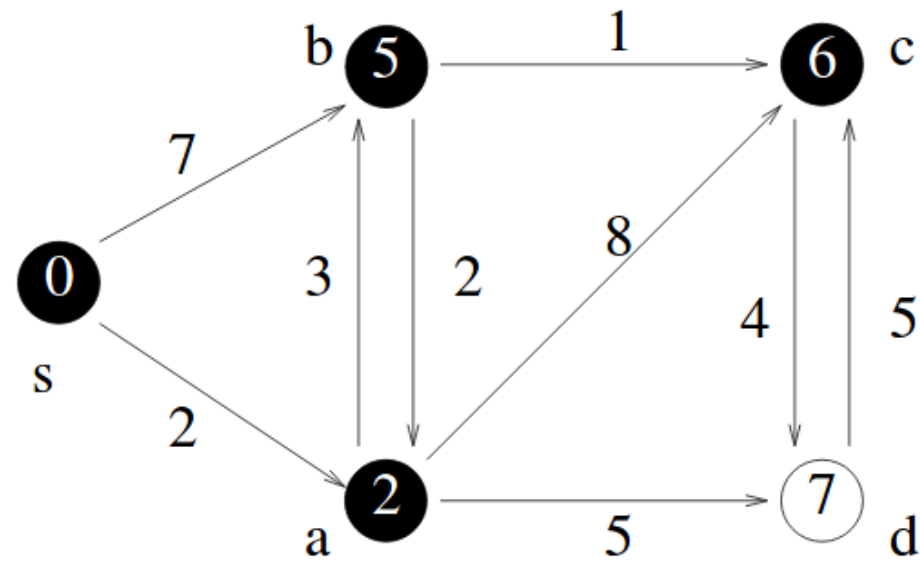


Example (cont.)

v	s	a	b	c	d
$d[v]$	0	2	5	6	7
$pred[v]$	nil	s	a	b	a
$color[v]$	B	B	B	W	W

v	c	d
$d[v]$	6	7

Example (cont.)

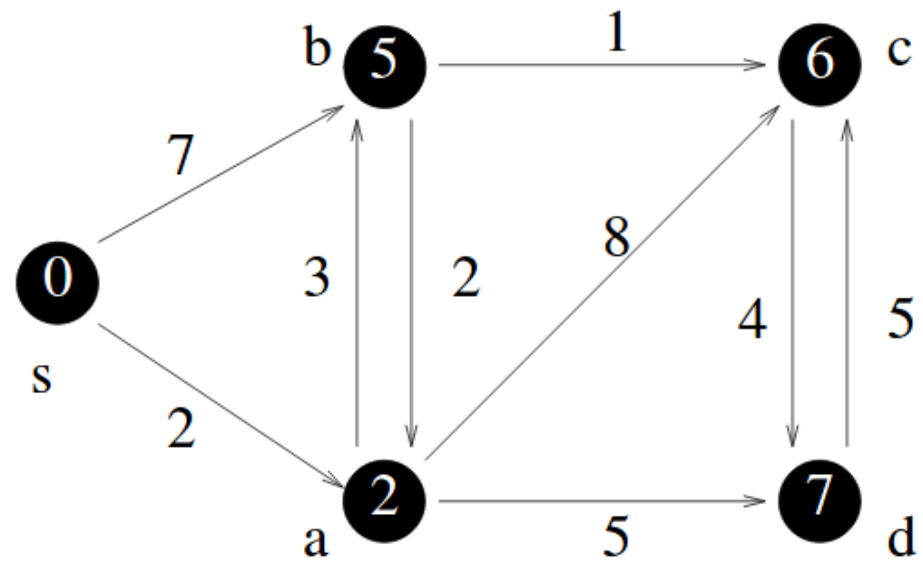


Example (cont.)

v	s	a	b	c	d
$d[v]$	0	2	5	6	7
$pred[v]$	nil	s	a	b	a
$color[v]$	B	B	B	B	W

v	d
$d[v]$	7

Example (cont.)



Example (cont.)

v	s	a	b	c	d
$d[v]$	0	2	5	6	7
$pred[v]$	nil	s	a	b	a
$color[v]$	B	B	B	B	B

$$Q = \emptyset.$$

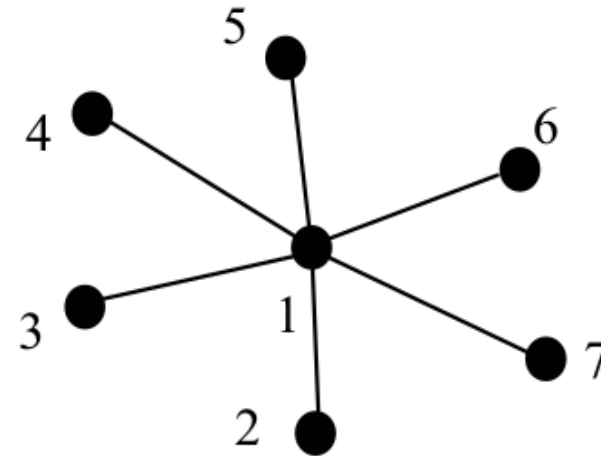
1.3 Degree Centrality/ Closeness Centrality

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

$d(i, j)$: shortest distance from node i to node j

Betweenness Centrality

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



$p_{jk}(i)$: Number of shortest path from node j to node k the pass node i

$$C_B(1) = 15, C_B(2) = C_B(3) = C_B(4) = C_B(5) = C_B(6) = C_B(7) = 0$$

1.4 Prestige/ Degree Prestige

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

$d_i(i)$: in-degree of node i

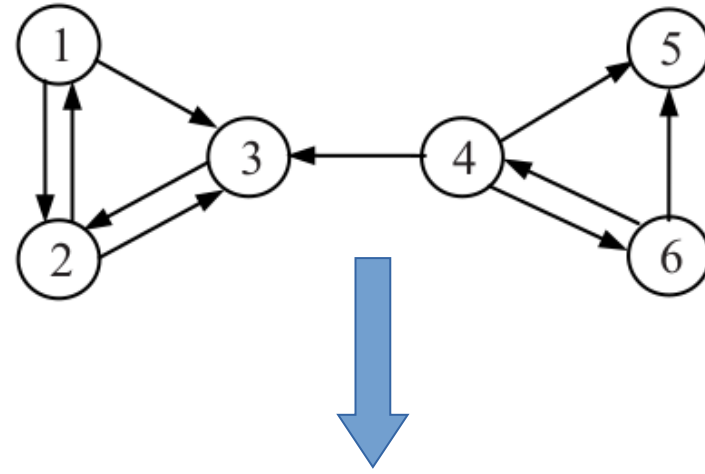
$$P_P(i) = \frac{|I_i|/(n-1)}{\sum_{j \in I_i} d(j,i) / |I_i|},$$

I_i : Set of nodes that can reach node i

1.5 PageRank Algorithm

- Rank graphs based on general structure
- For large graphs, the rank is approximated by an iterative algorithm based on the 'random walk'
- Important applications in web search engines
- Cons: Doesn't depend on the query

Transition matrix



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

Transition matrix normalization

Standardize:

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

$$R(A) = (1 - d) / N + d * \sum_{B:(B,A) \in E} R(B) / d_o(B)$$

$R(A)$: Thứ hạng của đỉnh A

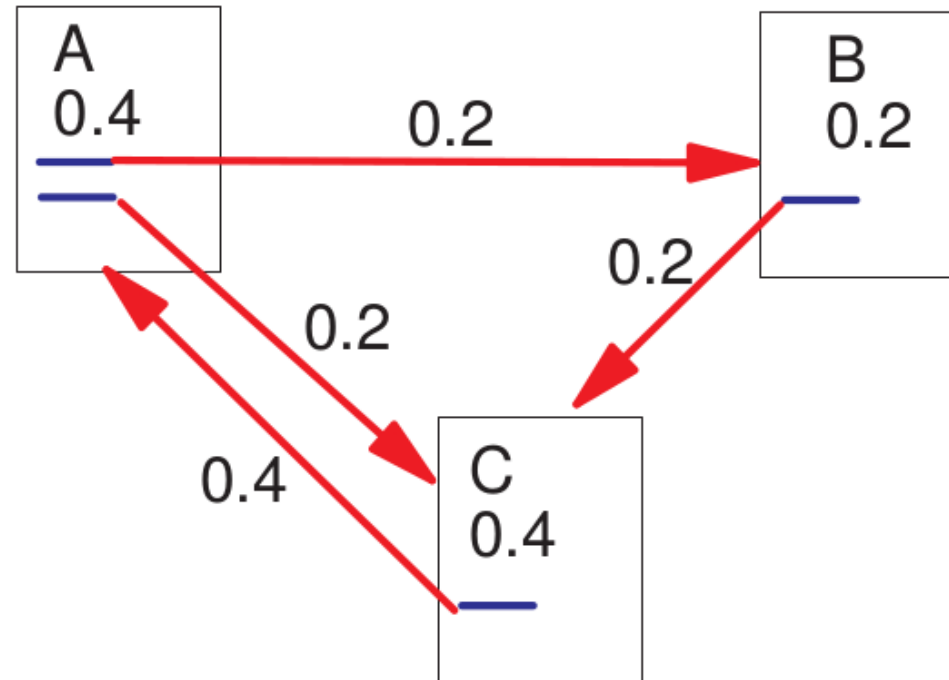
d : damping factor

N : số đỉnh của đồ thị

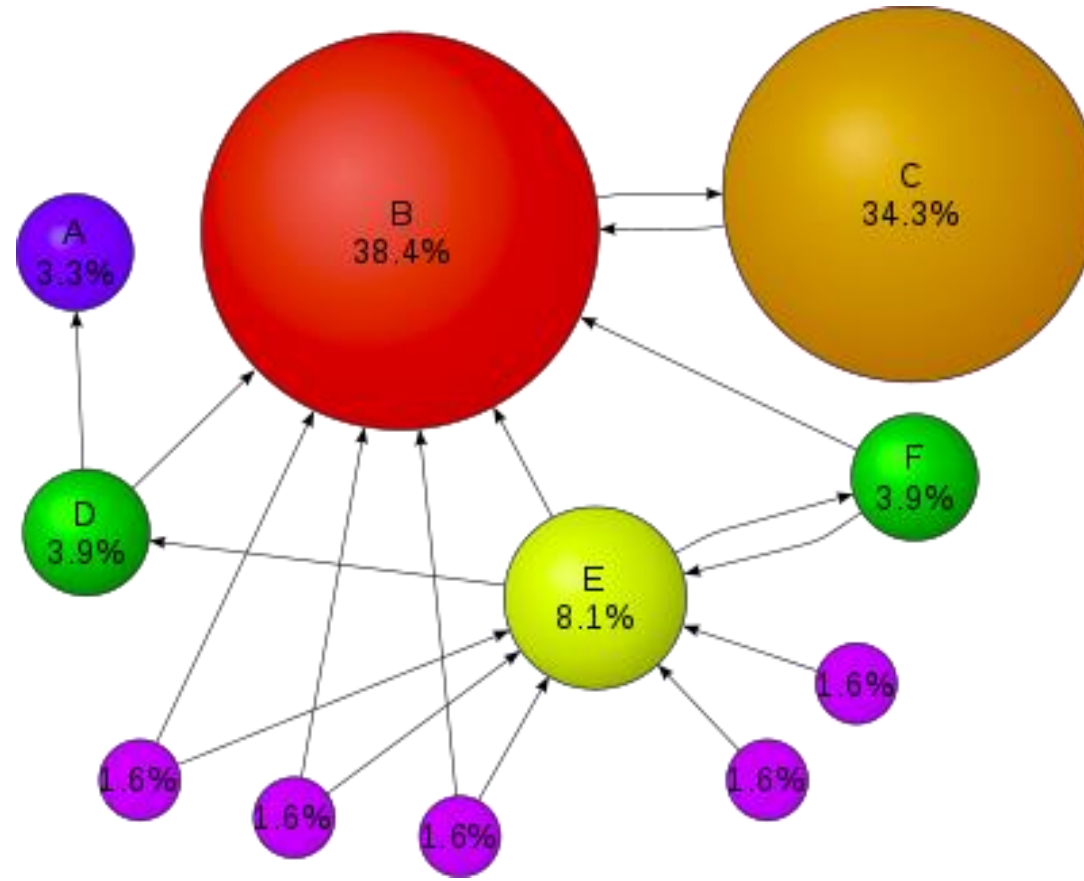
(B,A) cạnh của đồ thị

$d_o(B)$ bậc ra của đỉnh B

Example ($d=1$)



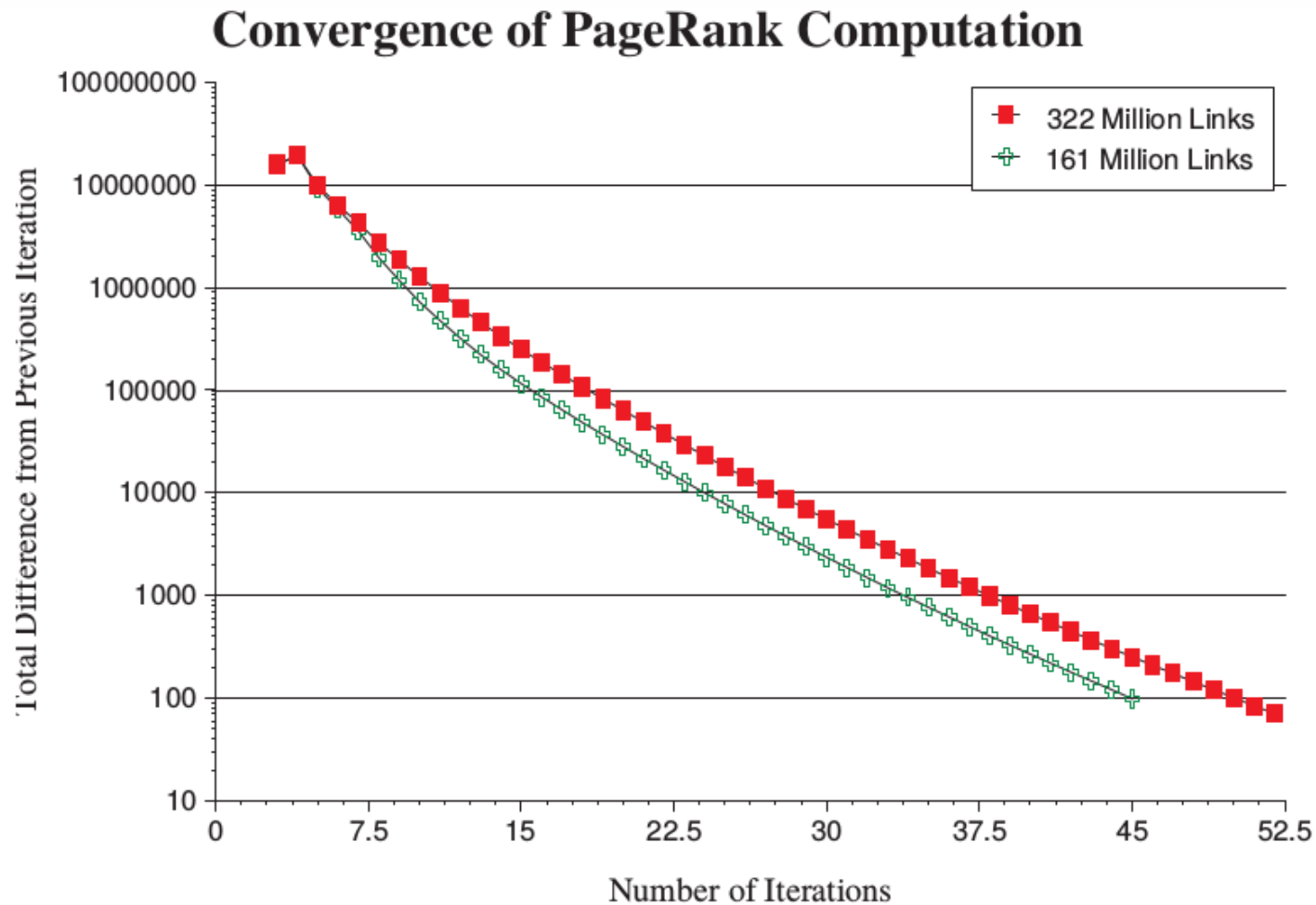
Example ($d = 0.85$)



Algorithm PageRank(d, E)

1. Init page ranks $R^{(0)}$;
2. $i = 1$;
3. **repeat**
4. **for** each page A **do**
5. $R^{(i)}(A) = (1 - d) / N + d * \sum_{B:(B,A) \in E} R^{(i-1)}(B) / d_o(B)$;
6. **endfor**
7. $i++$;
8. **until** converged

Convergence speed



Application: Web Search

Multi Search university [Next! \[national parks\]](#)

10 results clustering on Search

Query: **university**
11 Results Returned
Showing Results From 0 to 10

Stanford University Homepage
http://www.stanford.edu/
74.79% 4k - 2591993 - 010397

Stanford University Portfolio Collection
http://www.stanford.edu/home/administration/portfolio.html
65.78% 3k - 2591993 - 010397

University of Illinois at Urbana-Champaign
http://www.uiuc.edu/
73.26% 13k - 133096 - 010397

Indiana University
http://www.indiana.edu/
68.38% 1k - 092896 - 010597

University of California, Irvine
http://www.uci.edu/
68.07% 3k - 133096 - 010397

University of Minnesota
http://www.umn.edu/
67.05% 0k - 131696 - 010397

Iowa State University Homepage
http://www.iastate.edu/
66.66% 3k - 131896 - 010397

The University of Michigan
http://www.umich.edu/
66.35% 1k - 2591993 - 010397

Mississippi State University
http://www.msstate.edu/
66.35% 3k - 2591993 - 010397

Northwestern University: NUInfo
http://www.nwu.edu/
66.15% 3k - 131496 - 010597

next 10

Optical Physics at the University of Oregon
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group....
<http://optics.uoregon.edu/> - size 1K - 16 Dec 96

Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
<http://www.net.cmu.edu/> - size 4K - 19 Aug 95

Wesleyan University Computer Science Group Home Page
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
<http://www.cs.wesleyan.edu/> - size 2K - 15 Apr 96

Keio University Shonan Fujisawa Campus (SFC)
B\$3\$N%Z!EFnF#Bt%-%c%e%Q%99 (B(SFC) \$B\$N (BWWW \$B% \$BCmOU=q\$- (B \$B\$rF!\$s\$G\$!\$@!\$5\$!\$# (B. Nihongo | English. SFC \$B>pJs (B. [\$B%a%G%#%*%*%e%?!*...
<http://www.sfc.keio.ac.jp/> - size 3K - 5 Feb 97

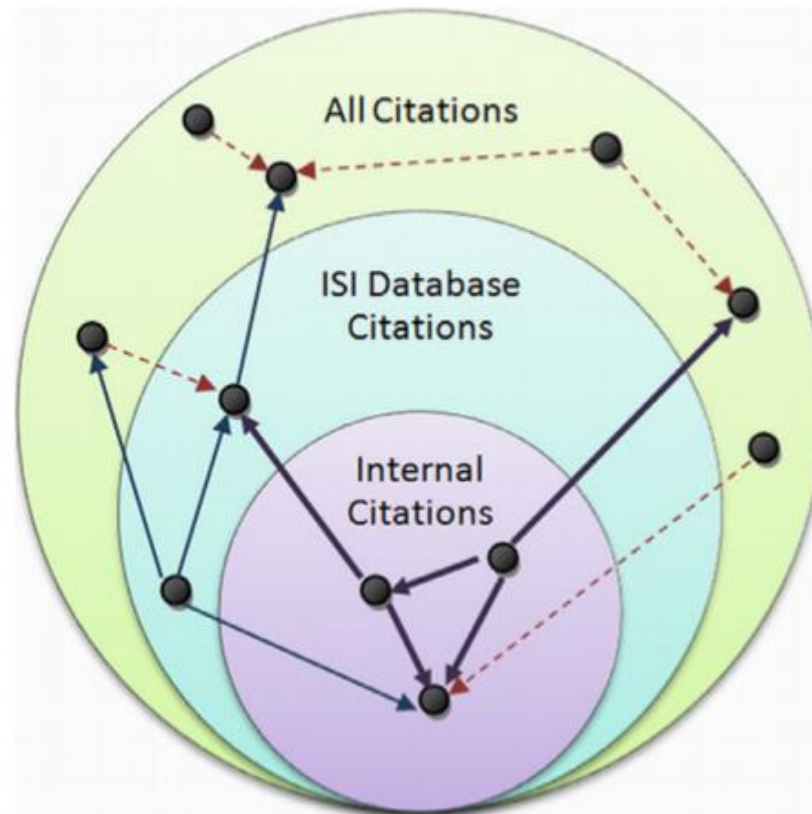
School of Chemistry, University of Sydney
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
<http://www.chem.su.oz.au/> - size 4K - 25 Feb 97

Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
<http://www.mankato.msus.edu/> - size 3K - 27 Nov 96

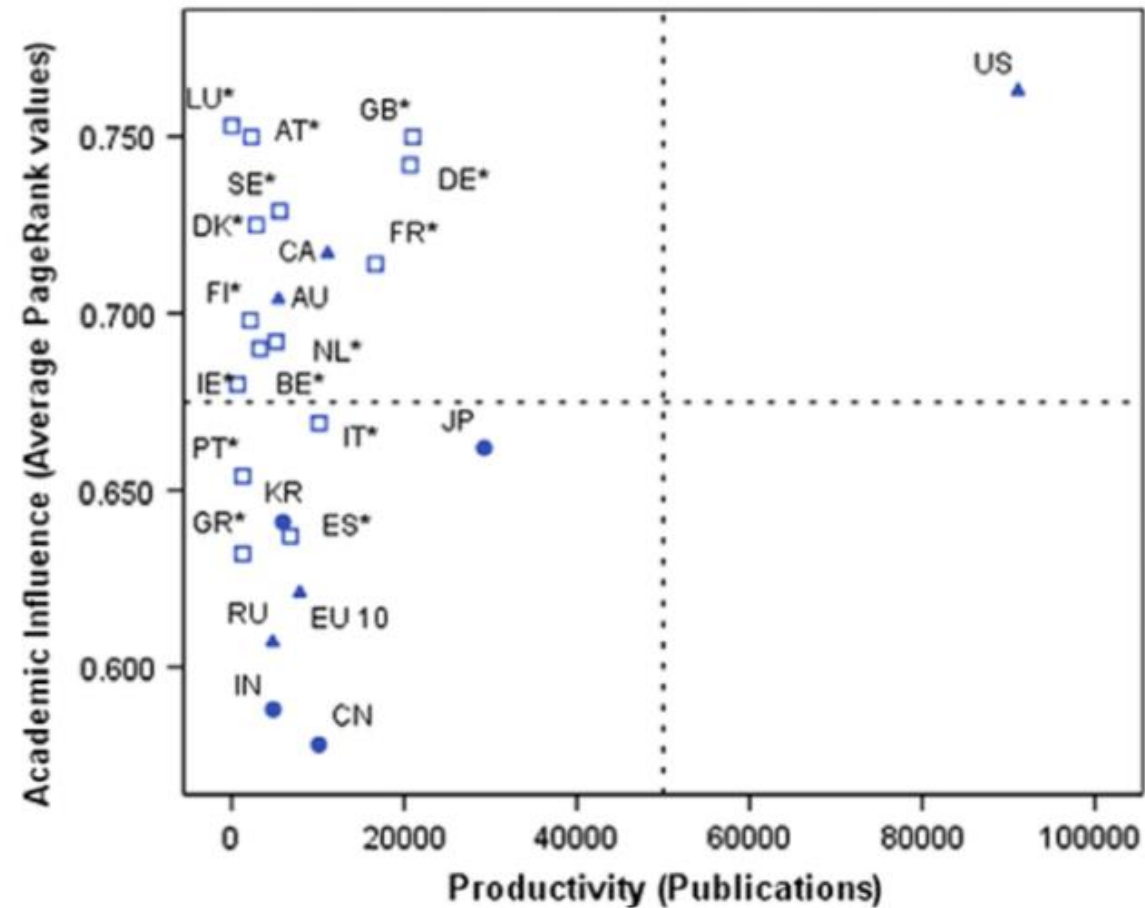
St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
<http://www.sau.edu/> - size 2K - 4 Feb 97

University of Washington ECSEL Projects

Guan et al. 2008. “*Bringing Page-Rank to the Citation Analysis*”







Application: Citation analysis (cont.)



1.6 HITS Algorithm

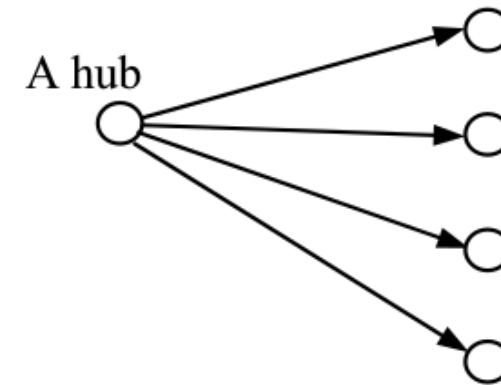
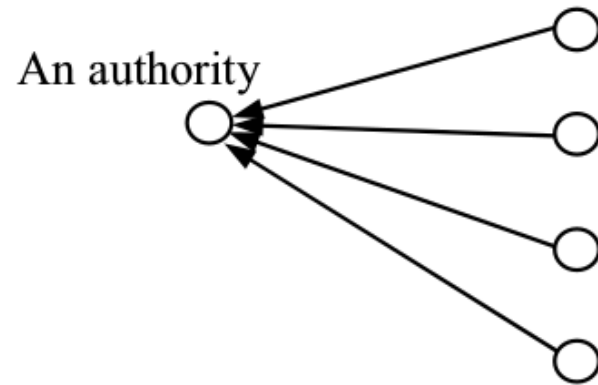
- Hypertext Induced Topic Search
- J. Kleinberg. “*Authoritative Sources in a Hyperlinked Environment.*” In Proc. of the 9th ACM SIAM Symposium on Discrete Algorithms (SODA’98), pp. 668–677, 1998.

	Spam filtering	Query relevance	Execution
HIST			Online
PageRank			Offline

Authority/Hub

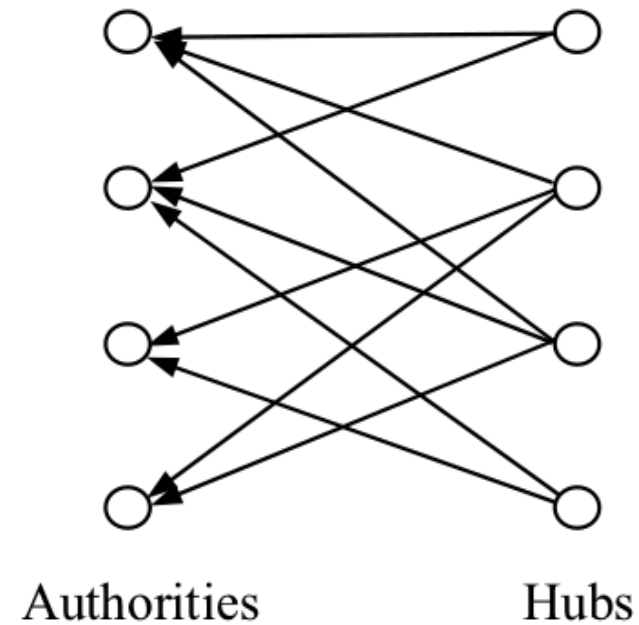
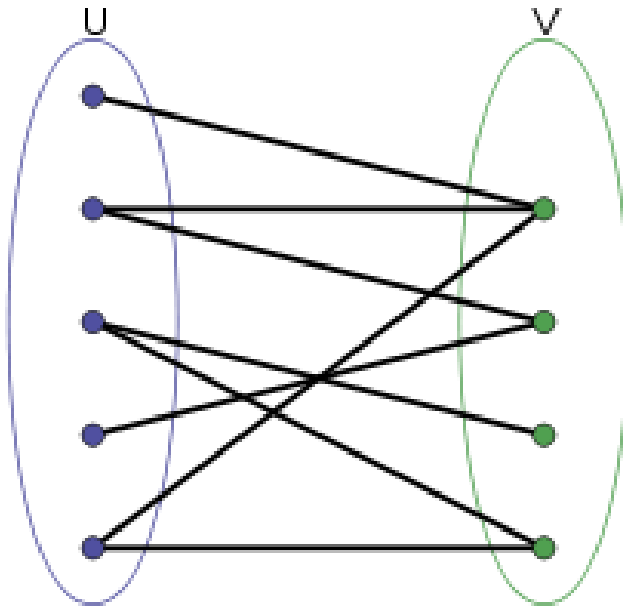
Authority: pages with many in-links

Hub: pages with many out-links



Bigraph

- Graph divided into 2 separated set of node such that every edge connects 2 node of different set



Algorithm

Input: Query q

Output: authority score and hub score of **relevant** pages of query q

Algorithm:

- 1 - *Retrieve information*
- 2 - *Expand graph*
- 3 - *Compute rank*

1-Retrieve information

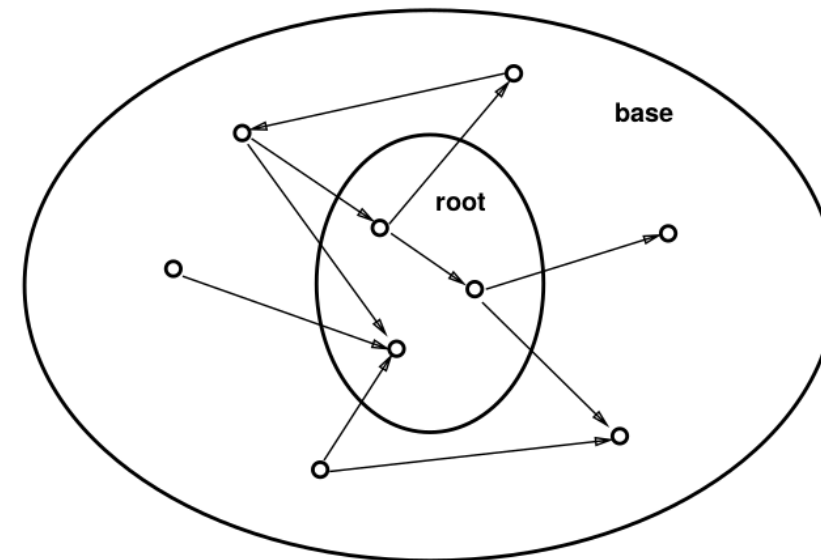
Requires a search engine has relevant documents of query q

- Input query q and a root set \mathbf{W} of top k pages relevant to q

2- Expand graph

From root set W , expand to base set S

- For each page p in W
 - Insert pages that p links to
 - Insert pages that links to p



3- Compute rank

Authority score (a)

Hub score (h)

$G = (V, E)$

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

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

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

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

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

3- Compute rank (cont.)

$$\mathbf{a} = \mathbf{L}^T \mathbf{h}$$

$$\mathbf{h} = \mathbf{L} \mathbf{a}$$

HITS-Iterate(G)

$\mathbf{a}_0 \leftarrow \mathbf{h}_0 \leftarrow (1, 1, \dots, 1);$

$k \leftarrow 1$

Repeat

$\mathbf{a}_k \leftarrow \mathbf{L}^T \mathbf{L} \mathbf{a}_{k-1};$

$\mathbf{h}_k \leftarrow \mathbf{L} \mathbf{L}^T \mathbf{h}_{k-1};$

$\mathbf{a}_k \leftarrow \mathbf{a}_k / \|\mathbf{a}_k\|_1; \quad // \text{normalization}$

$\mathbf{h}_k \leftarrow \mathbf{h}_k / \|\mathbf{h}_k\|_1; \quad // \text{normalization}$

$k \leftarrow k + 1;$

until $\|\mathbf{a}_k - \mathbf{a}_{k-1}\|_1 < \varepsilon_a$ and $\|\mathbf{h}_k - \mathbf{h}_{k-1}\|_1 < \varepsilon_h;$

return \mathbf{a}_k and \mathbf{h}_k

A large graphic on the left side of the slide. It features a dark blue background with a circular pattern of red dots of varying sizes, creating a sense of depth and movement. The word "HUST" is centered within this graphic in a white, bold, sans-serif font.

HUST

THANK YOU !