



ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

# BÀI 5: PHÂN TÍCH LIÊN KẾT

# Các bài toán chính trong phân tích liên kết

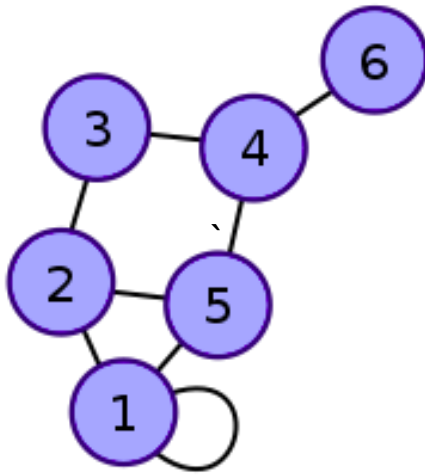
- Xếp hạng đồ thị: Phân tích vai trò của các đỉnh trong đồ thị
- Nhận diện cộng đồng: Phát hiện các cộng đồng bao gồm các thành viên có tính chất tương tự
- Dự đoán liên kết: Dự đoán sự tiến hóa của đồ thị theo thời gian
- Phân loại đồ thị: Phân loại các đỉnh và các cạnh của đồ thị vào các lớp cho trước

# Nội dung

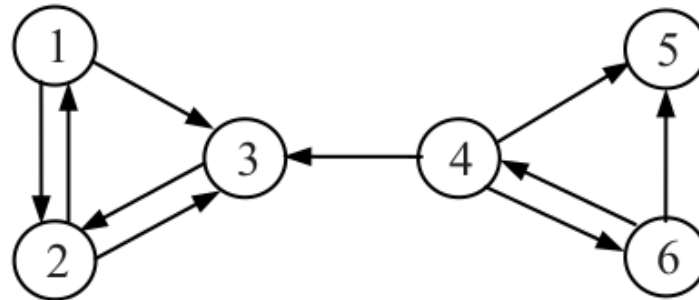
1. Xếp hạng đồ thị
2. Nhận diện cộng đồng
3. Học biểu diễn đồ thị

# 1. Xếp hạng đồ thị

## 1.1 Các khái niệm cơ bản của đồ thị



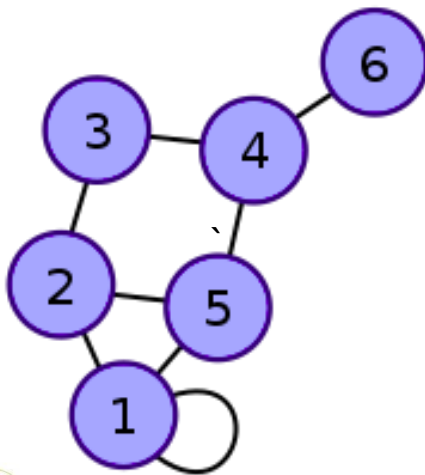
a) Đồ thị vô hướng



b) Đồ thị có hướng

# Ma trận kề

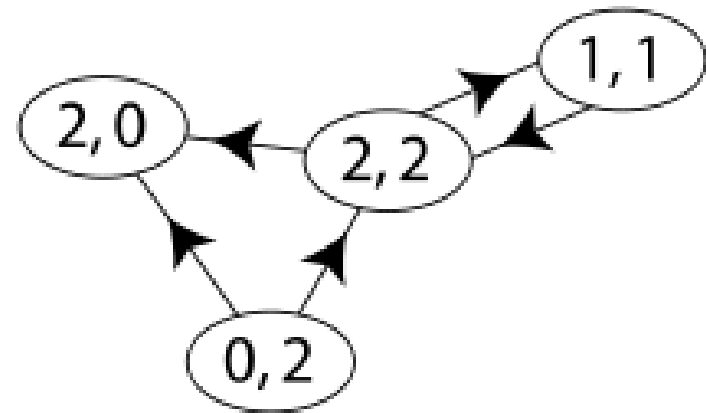
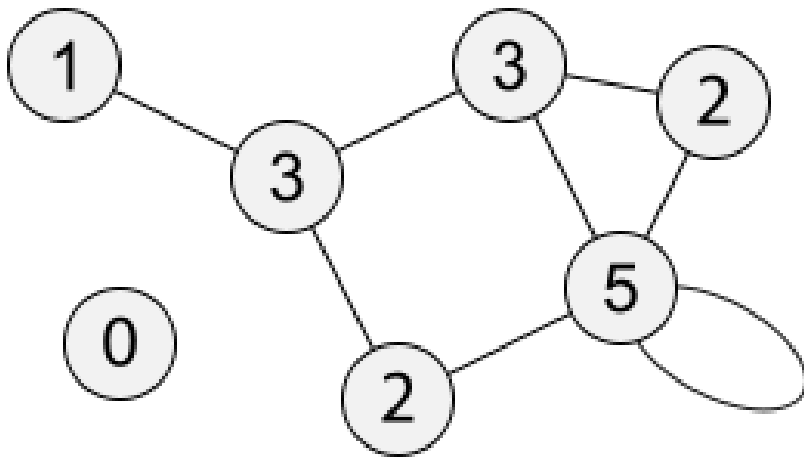
$$a[i, j] \begin{cases} = 1 \text{ nếu tồn tại cạnh } (i, j) \\ = 0 \text{ nếu ngược lại} \\ = 2 \text{ nếu tồn tại cạnh từ một đỉnh đến chính nó} \end{cases}$$



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

# Bậc của đỉnh

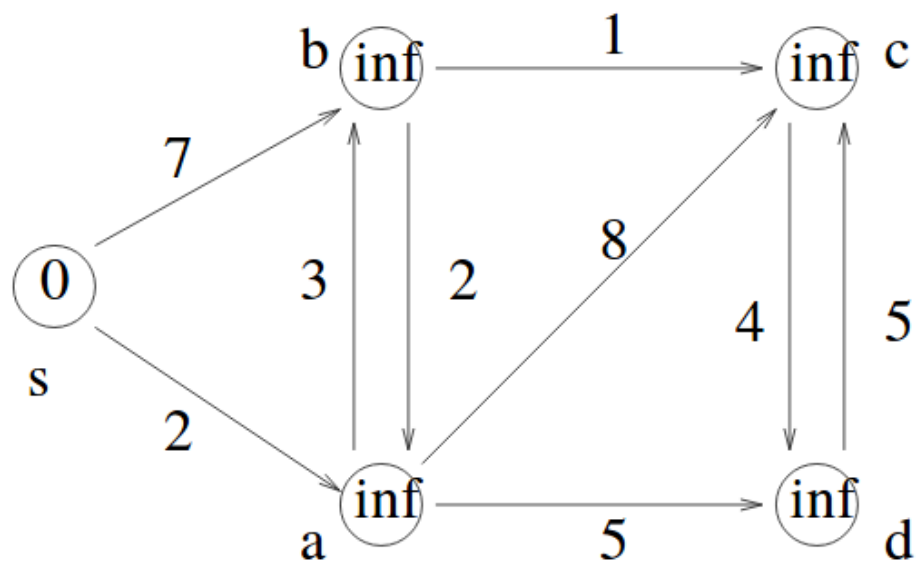
- $d_i(i) =$  số nút trở tới  $i$
- $d_o(i) =$  số nút  $i$  trở tới



# 1.2 Thuật toán Dijkstra

- Tìm đường đi ngắn nhất từ một đỉnh  $s$  tới các đỉnh còn lại của đồ thị
  - $d(v)$ : Khoảng cách từ đỉnh  $v$  tới đỉnh  $s$ 
    - B1**: Khởi tạo  $d(s) = 0$ ;  $d(v) = \infty$
    - B2**: Sắp xếp các đỉnh  $v$  theo một trật tự xác định trên hàng đợi  $Q$
    - B3**: Lấy một đỉnh  $u$  thuộc hàng đợi  $Q$  và cập nhật khoảng cách  $d(v)$  (nếu cần) với mỗi đỉnh  $v$  liền kề với  $u$
- Quay lại **B2** cho đến khi xử lý hết các đỉnh

# VD



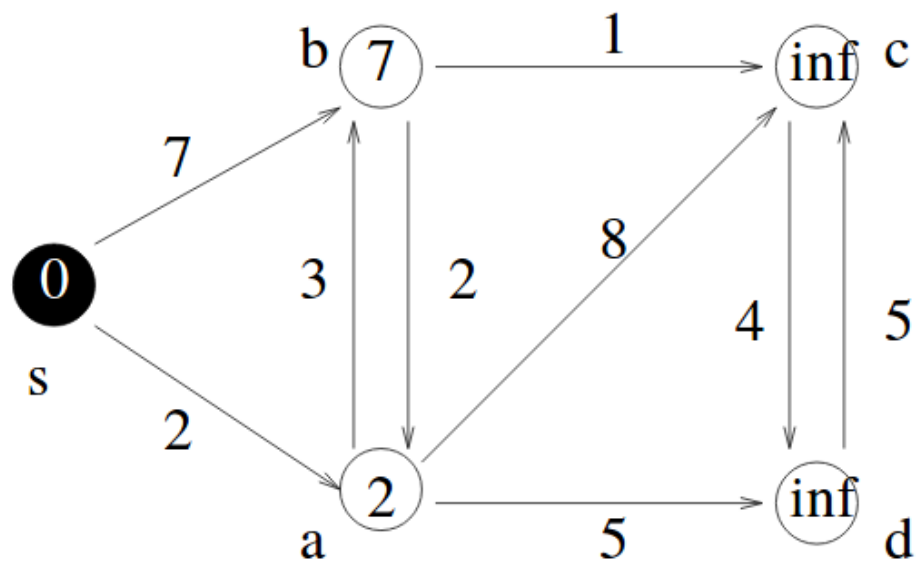


# VD (tiếp)

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

$v$	s	a	b	c	d
$d[v]$	0	$\infty$	$\infty$	$\infty$	$\infty$

# VD (tiếp)

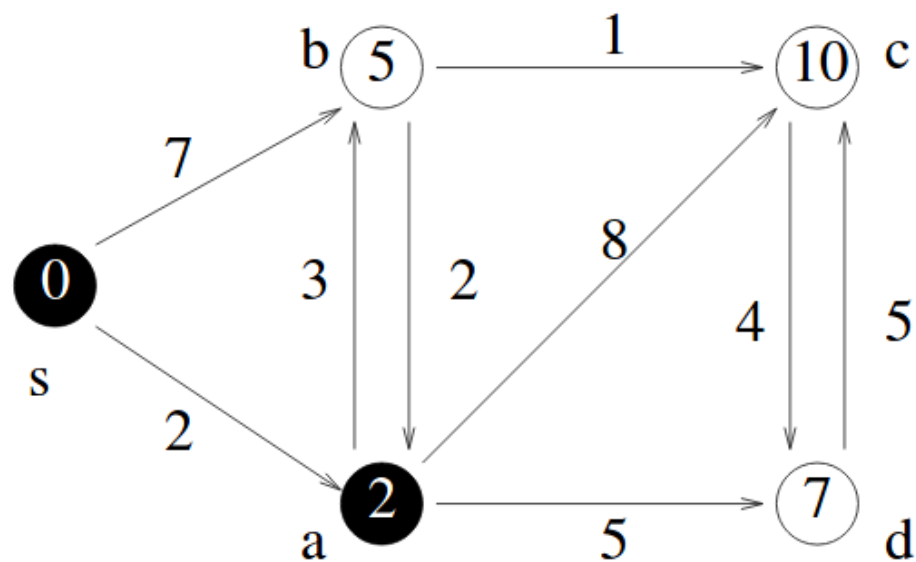


# VD (tiếp)

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

$v$	a	b	c	d
$d[v]$	2	7	$\infty$	$\infty$

# VD (tiếp)

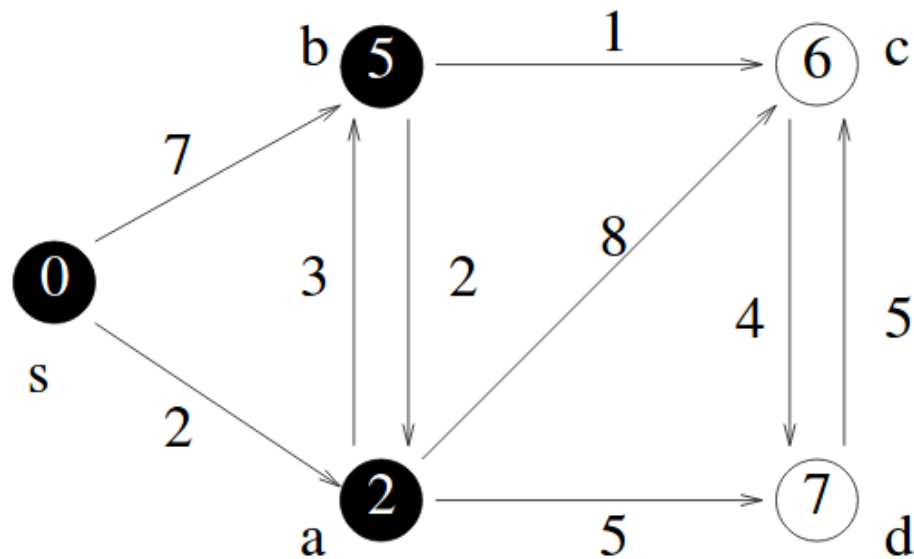


# VD (tiếp)

$v$	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

# VD (tiếp)

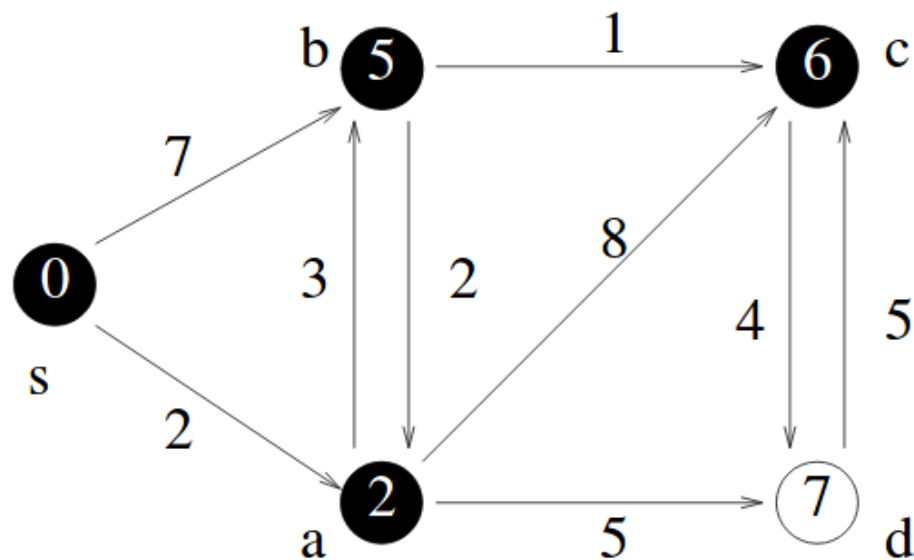


# VD (tiếp)

$v$	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

# VD (tiếp)



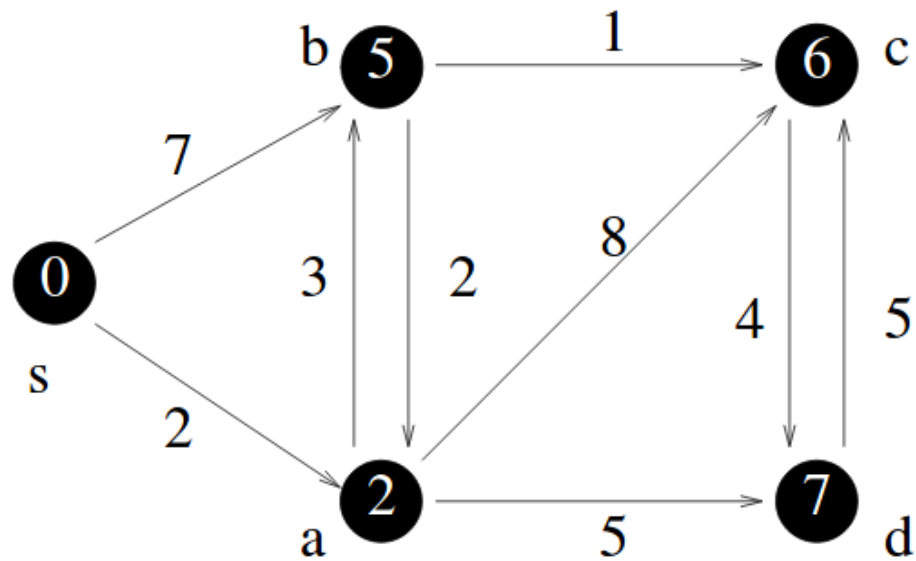


# VD (tiếp)

$v$	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

# VD (tiếp)



# VD (tiếp)

$v$	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 Độ trung tâm

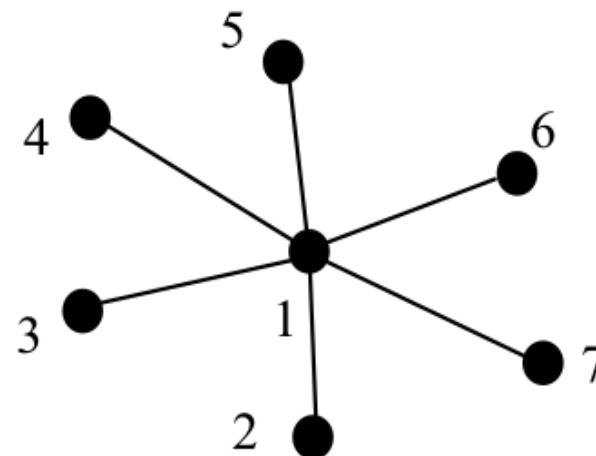
## Độ trung tâm lân cận

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

$d(i, j)$ : Khoảng cách ngắn nhất từ nút  $i$  tới nút  $j$

# Độ trung tâm trung gian

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



$p_{jk}(i)$ : Số lượng đường đi ngắn nhất từ  $j$  tới  $k$  mà đi qua  $i$

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

# 1.4 Độ quan trọng

## Độ quan trọng theo bậc

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

$d_i(i)$ : Số nút trở tới  $i$

# Độ quan trọng lân cận

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

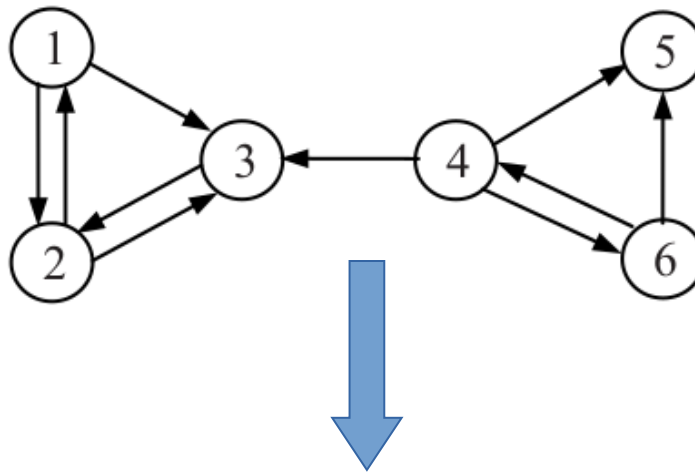
$I_i$ : Các nút có thể đi tới  $i$

# 1.5 Thuật toán Pagerank

- Xếp hạng đồ thị dựa trên cấu trúc tổng quát
- Đối với các đồ thị lớn, thứ hạng được tính xấp xỉ bằng thuật toán lặp dựa trên '*random walk*'
- Có ứng dụng quan trọng trong máy tìm kiếm web
- Nhược điểm: Không phụ thuộc vào câu truy vấn



# Ma trận chuyển tiếp



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

# Ma trận chuyển tiếp (tiếp)

Chuẩn hóa:

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

# Công thức xếp hạng

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

$R(A)$ : Thứ hạng của đỉnh A

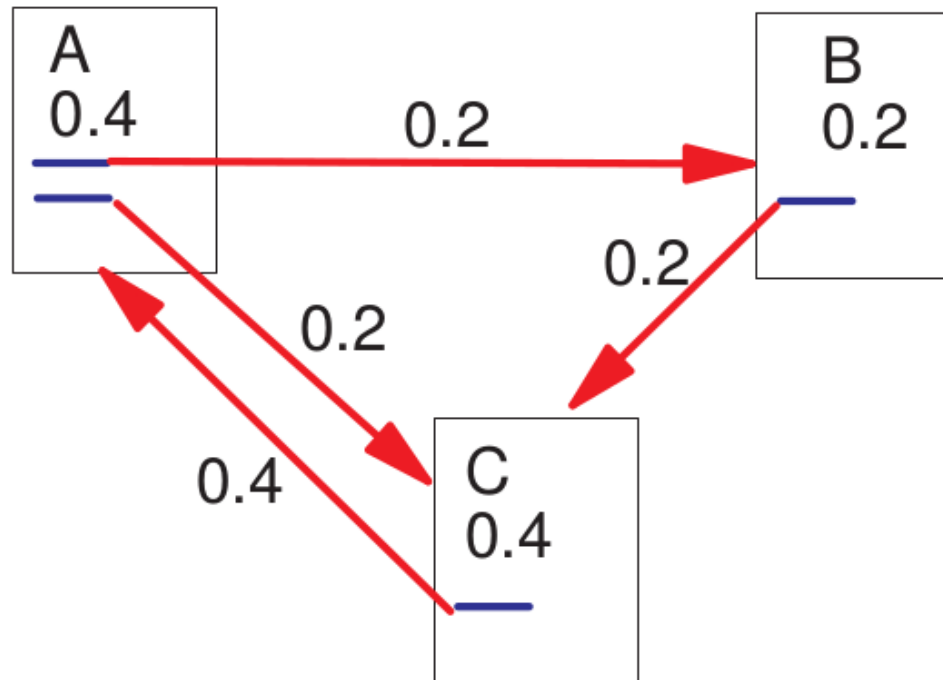
$d$ : damping factor

$N$ : số đỉnh của đồ thị

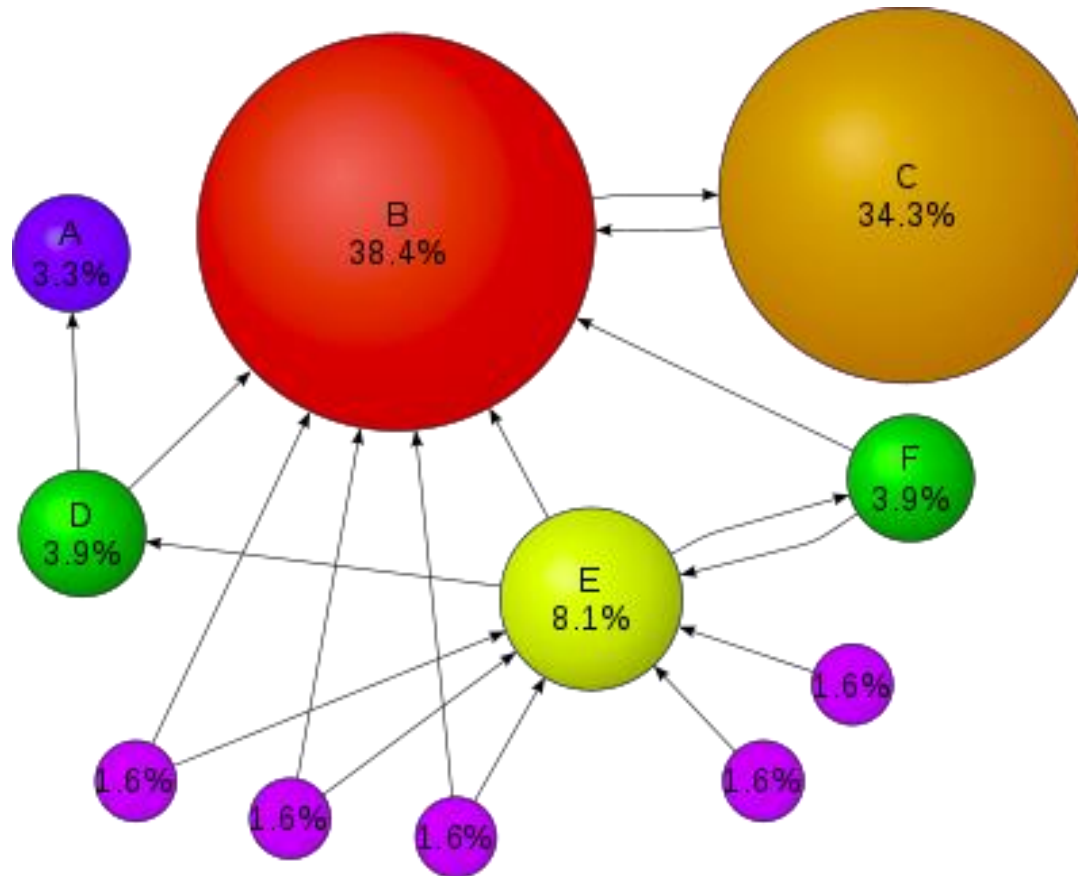
$(B,A)$  cạnh của đồ thị

$d_o(B)$  bậc ra của đỉnh B

# VD ( $d = 1$ )



# VD ( $d = 0.85$ )

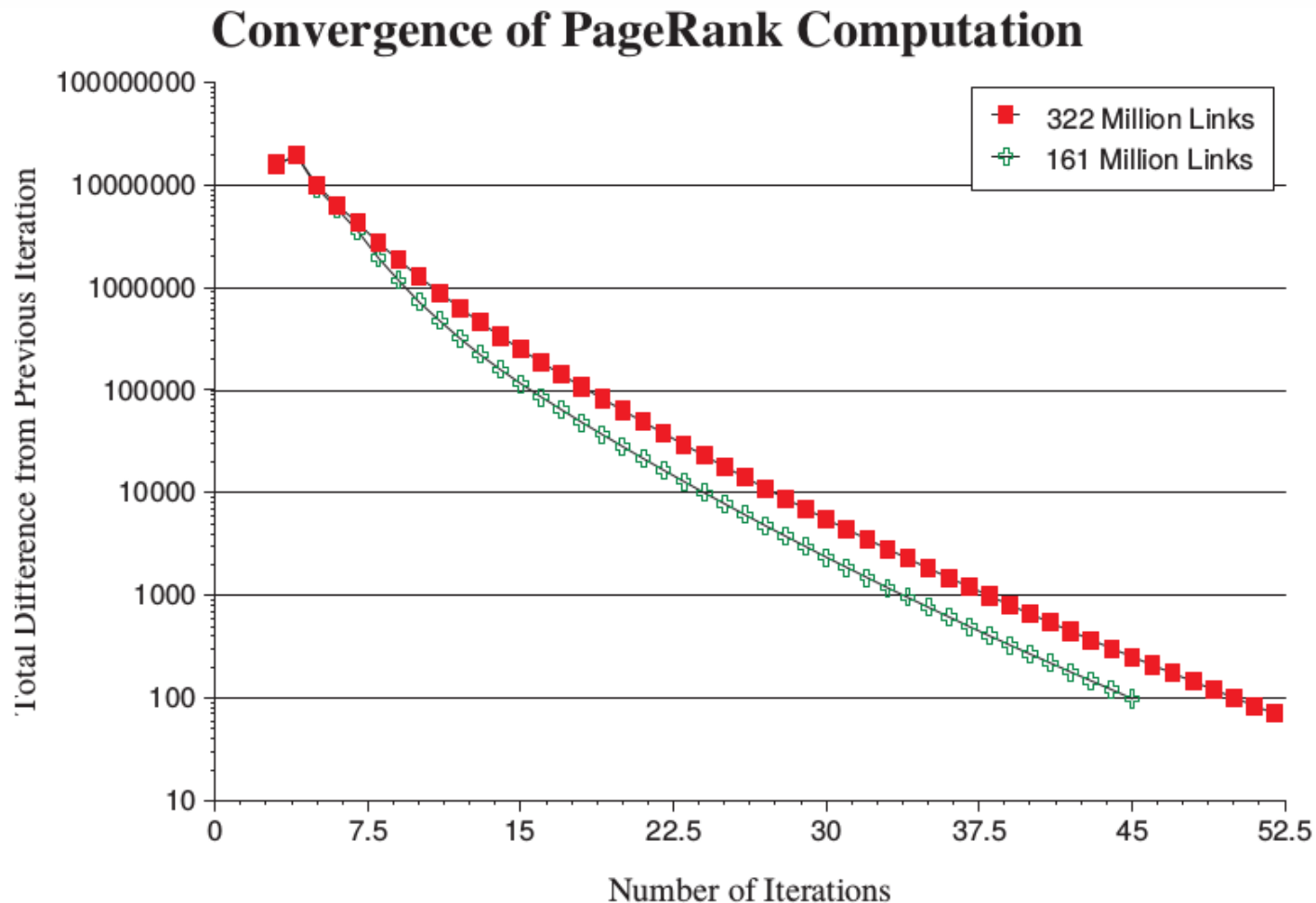


# Thuật toán lặp

## Algorithm PageRank( $d, E$ )

1. Khởi tạo thứ hạng các trang  $R^{(0)}$ ;
2.  $i = 1$ ;
3. **repeat**
4.     **for** mỗi trang 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** hội tụ

# Tốc độ hội tụ



# Ứng dụng: Tìm kiếm Web

The screenshot displays a web search interface with a search bar at the top containing the text "Multi Search" and "university". To the right of the search bar is a "Search" button and a link "Next! [national parks]". Below the search bar, there are tabs for "10 results", "clustering on", and "Search". The main content area is divided into two columns. The left column lists search results for the query "university", showing 11 results returned. The results are listed with their titles, URLs, and some additional information like file size and date. The right column displays the content of the selected result, "Optical Physics at the University of Oregon".

Multi Search university Search 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 - 3/5/1993 - 01/03/97

**Stanford University Portfolio Collection**  
http://www.stanford.edu/home/administration/portfolio.html  
65.78% 3K - 3/5/1993 - 01/03/97

**University of Illinois at Urbana-Champaign**  
http://www.uiuc.edu/  
73.26% 13K - 12/30/96 - 01/03/97

**Indiana University**  
http://www.indiana.edu/  
68.38% 1K - 09/28/96 - 01/05/97

**University of California, Irvine**  
http://www.uci.edu/  
68.07% 2K - 12/30/96 - 01/03/97

**University of Minnesota**  
http://www.umn.edu/  
67.05% 0K - 12/16/96 - 01/03/97

**Iowa State University Homepage**  
http://www.iastate.edu/  
66.66% 3K - 12/18/96 - 01/03/97

**The University of Michigan**  
http://www.umich.edu/  
66.35% 1K - 3/5/1993 - 01/03/97

**Mississippi State University**  
http://www.msstate.edu/  
66.35% 3K - 3/5/1993 - 01/03/97

**Northwestern University NUIInfo**  
http://www.nwu.edu/  
66.15% 3K - 12/14/96 - 01/05/97

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\$\$\$N%ZIEFnF#Bt%-9c%9s%Q%99 (B(SFC) \$B\$N (BWWW \$B% \$BCmOU=q%- (B \$B\$rFI\$s\$G\$%\$@%\$%\$!# (B. Nihongo | English. SFC \$B>pJs (B. [ \$B%a%G%#%\*%9%9s%?!\*...  
<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.msut.edu/> - size 3K - 27 Nov 96

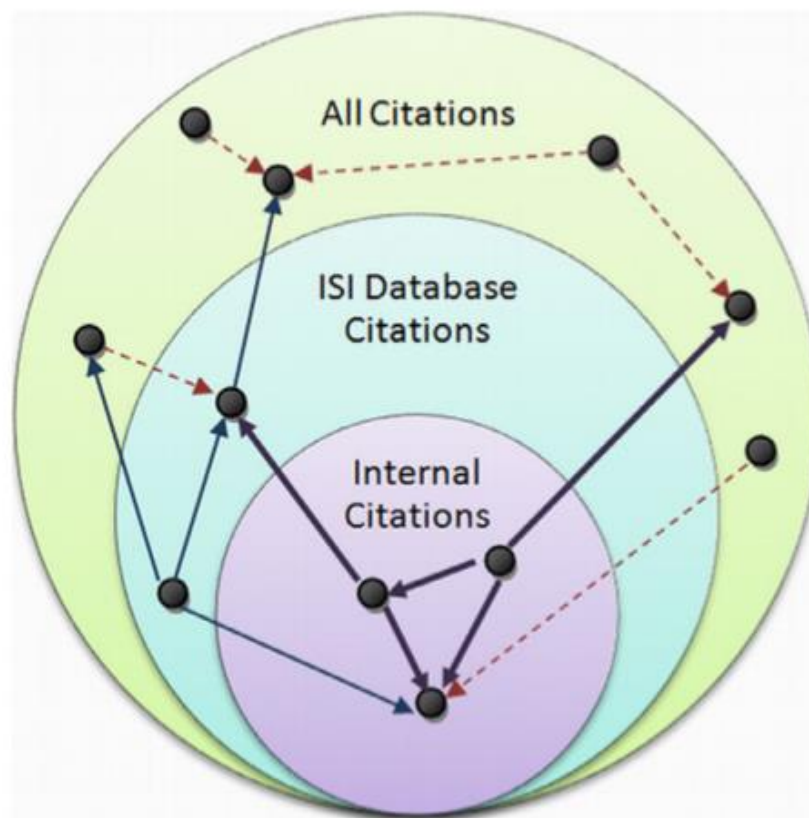
**St. Ambrose University**  
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...  
<http://www.sau.edu/> - size 2K - 4 Feb 97

**University of Washington ECSEL Projects**

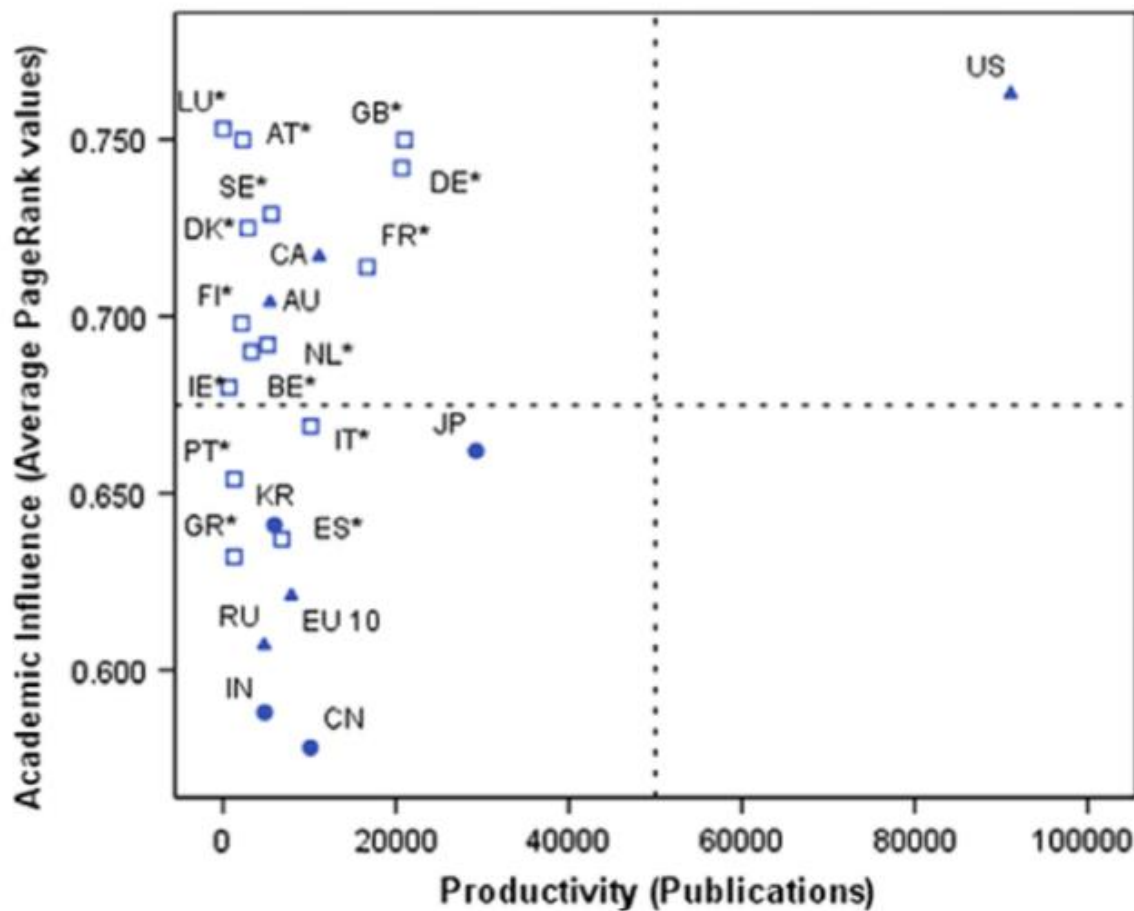


# Ứng dụng: Phân tích trích dẫn

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







# Ứng dụng: Phân tích trích dẫn (tiếp)



## •1.6 Thuật toán HITS

# 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.

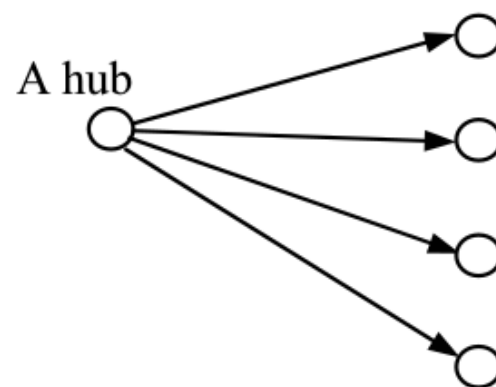
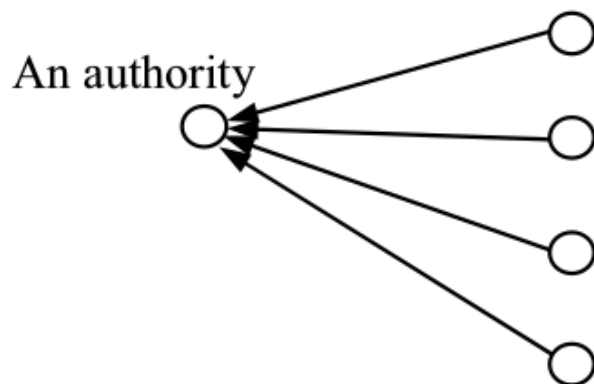
	Spam filtering	Query relevance	Execution
HIST			Online
PageRank			Offline

# Authority/Hub

Authority: Trang được trỏ tới nhiều

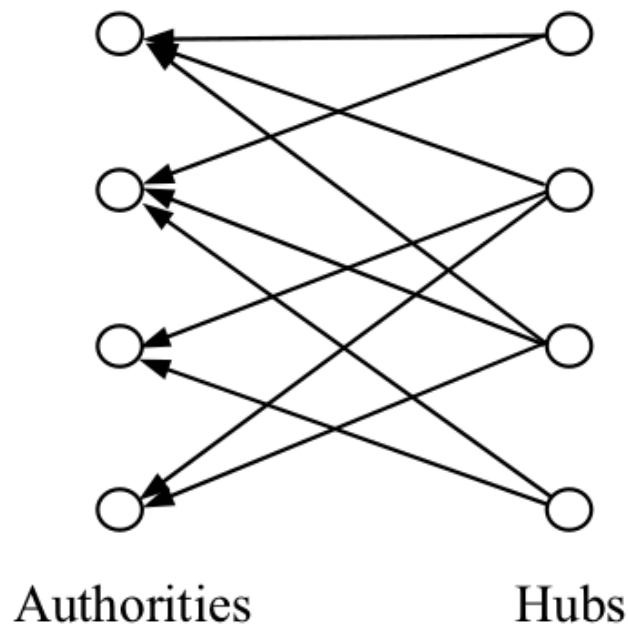
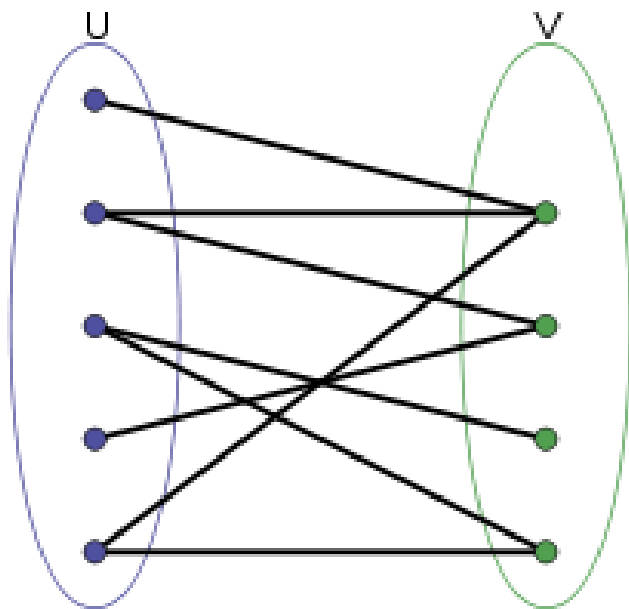
Hub: Trang trỏ tới nhiều trang khác

Authority và hub có mối quan hệ tương hỗ



# Bigraph

- Các nút chia thành hai tập không giao nhau
- Mỗi cạnh đều nối hai nút thuộc hai tập



# Thuật toán

Đầu vào: *Câu truy vấn  $q$*

Đầu ra: Điểm authority và hub của các trang **liên quan** đến  $q$

Thuật toán:

1 - *Truy hồi thông tin*

2 - *Mở rộng đồ thị*

3 - *Tính ranking*

# 1-Truy hồi thông tin

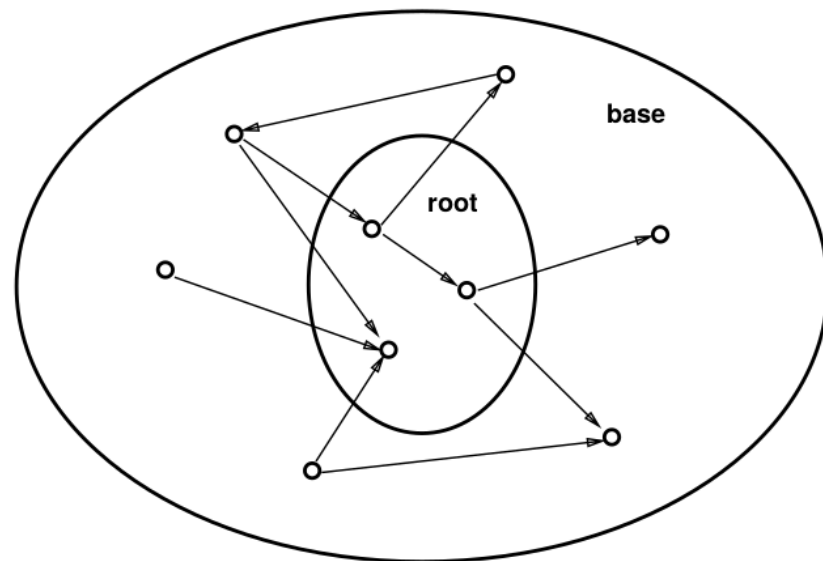
Y/c một máy tìm kiếm có chứa các văn bản liên quan đến câu truy vấn  $q$  (vd Google, Coccoc)

- Đưa  $q$  vào máy tìm kiếm và lấy về tập root  $\mathbf{W}$  gồm  $k$  trang liên quan nhất đến  $q$  (vd  $k = 200$ )

## 2- Mở rộng đồ thị

Từ tập root  $W$ , mở rộng ra tập base  $S$

- Với mỗi trang  $p$  trong  $W$ 
  - Bổ sung các trang mà  $p$  trở tới
  - Bổ sung các trang trở tới  $p$





# 3- Tính thứ hạng

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- Tính thứ hạng (tiếp)

$$a = L^T h$$

$$h = La$$

**HITS-Iterate( $G$ )**

$a_0 \leftarrow h_0 \leftarrow (1, 1, \dots, 1);$

$k \leftarrow 1$

**Repeat**

$a_k \leftarrow L^T L a_{k-1};$

$h_k \leftarrow L L^T h_{k-1};$

$a_k \leftarrow a_k / \|a_k\|_1; \quad // \text{normalization}$

$h_k \leftarrow h_k / \|h_k\|_1; \quad // \text{normalization}$

$k \leftarrow k + 1;$

**until**  $\|a_k - a_{k-1}\|_1 < \varepsilon_a$  and  $\|h_k - h_{k-1}\|_1 < \varepsilon_h;$

**return**  $a_k$  and  $h_k$



25 YEARS ANNIVERSARY  
**SOICT**

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG  
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

**Thank you for  
your attentions!**



[soict.hust.edu.vn/](http://soict.hust.edu.vn/)



[fb.com/groups/soict](https://fb.com/groups/soict)

