



Link Analysis and Web Search



Outline

- Ranking of web pages
- Algorithms that rank web pages
 - Hubs and Authorities
 - HITS (hyperlink-induced topic search), ask.com
 - Authoritative sources in a hyperlinked environment, Jon Kleinberg, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
 - PageRank
 - The anatomy of a large-scale hypertextual web search engine, Sergey Brin and Lawrence Page, Computer Networks and ISDN Systems, vol 30, Issue 1-7, 1998, p 107-117



The Problem of Ranking

- How does the search engine know which is the best answer ?
 - From 1994 to now
 - 1994 : 1500 queries per day
 - 1997 : Altavista : 20 million queries per day
 - Scaling with the web
 - Avoid junk results



Google search results for "universities hong kong". The search bar shows the query and a microphone icon. Below the search bar are tabs for All, Maps, Images, News, Videos, More, Settings, and Tools. The results section is titled "Hong Kong > Colleges and Universities" and displays a grid of university logos and names, including The University of Hong Kong, Chinese University of Hong Kong, City University of Hong Kong, Hong Kong Polytechnic University, Hong Kong University of Science and Technology, The Open University of Hong Kong, and The Education University of Hong Kong.

List of higher education institutions in Hong Kong - Wikipedia

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

Note 3: In August 2017, Hong Kong government announced 6 self-funded institutions, Caritas Institute of Higher Education, Chu Hai College of Higher Education, Hang Seng Management College, The Open University of Hong Kong, Tung Wah College and Technological and Higher Education Institute of Hong Kong will be included ...

[UGC-funded universities](#) · [Self-funded institutions](#) · [Sub-degree institutions](#)

The Chinese University of Hong Kong

www.cuhk.edu.hk/

The Chinese University of Hong Kong (CUHK) is a top Hong Kong university with strong research emphasis. The university aims to bring together China and the West.

The University of Hong Kong (HKU)

<https://www.hku.hk/>

Established in 1911, the University of Hong Kong (HKU) is the territory's oldest institute of higher learning and also an internationally recognized, research led, comprehensive university.

[Faculties & Departments](#) · [HKU Portal](#) · [Undergraduate Admissions](#) · [Contact Us](#)

All universities in Hongkong - Hotcourses Abroad

<https://www.hotcoursesabroad.com> > [Asia](#) > [Hong Kong](#) > [Universities List](#)

Map of Hong Kong and surrounding areas (Shenzhen, Zhuhai, Macau). The map shows the location of Hong Kong relative to the mainland and other regions.

Hong Kong

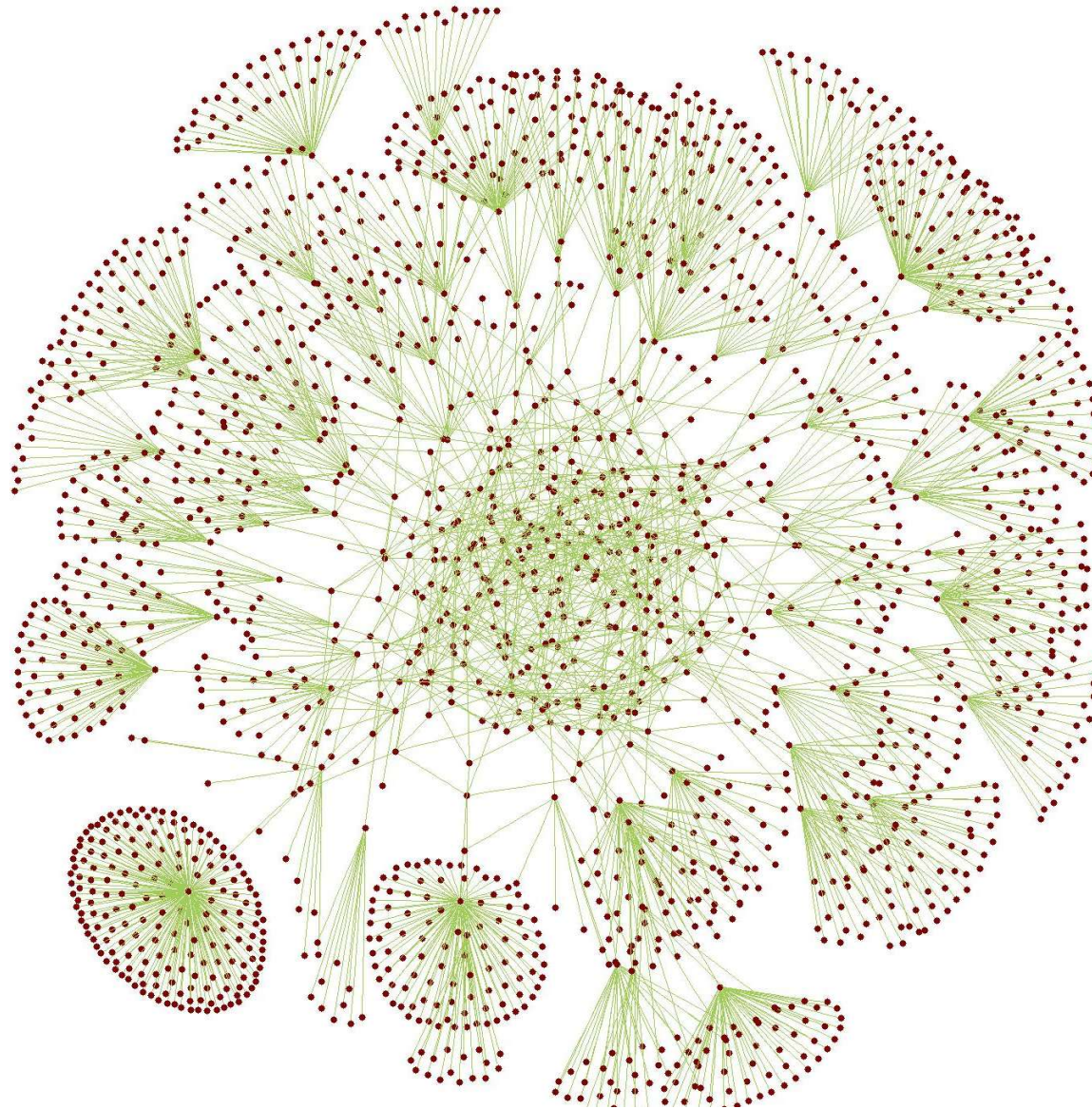
Chinese special administrative region

Hong Kong is an autonomous territory, and former British colony, in southeastern China. Its vibrant, densely populated urban centre is a major port and global financial hub with a skyscraper-studded skyline. Central (the business district) features architectural landmarks like I.M. Pei's Bank of China Tower. Hong Kong is also a major shopping destination, famed for bespoke tailors and Temple Street Night Market.



The Problem of Ranking

- How does the search engine know which is the best answer ?
- Search engines rank pages using automatic methods based on information intrinsic to the Web and its structure.





Information Retrieval

- Started from 1960s
- Search repositories of newspaper articles, scientific papers, patents, legal abstracts, and other document collections in response to keyword queries.
- Now : web



Information Retrieval

- Hard problem
 - synonymy
 - multiple ways to say the same thing
 - E.g. scallions and green onions
 - polysemy : multiple meanings for the same term
 - jaguar
 - animal
 - automobiles,
 - American football team in Florida,
 - an operating system for the Apple Macintosh





Apple Introduces “Jaguar,” the Next Major Release of Mac OS X

Version 10.2 Has More Than 150 New Features & Applications

MACWORLD EXPO, NEW YORK—July 17, 2002—Apple® today introduced Mac® OS X version 10.2 “Jaguar,” the next major release of Mac OS X featuring more than 150 amazing new features and applications. “Jaguar” includes a new Mail application designed to eliminate junk mail, iChat AIM-compatible instant messenger, a system-wide Address Book, Inkwell™ handwriting recognition, QuickTime® 6 with MPEG-4, improved Universal Access, an enhanced Finder™, Sherlock® 3 with Internet Services and Rendezvous™, Apple’s revolutionary home networking technology. Mac OS X v10.2 “Jaguar” will be publicly available August 24 for a suggested retail price of \$129 (US).


“Jaguar is light years ahead of Windows XP. There’s never been a better time to switch to Mac,” said Steve Jobs, Apple’s CEO. “With Unix at its core, and the most advanced object-oriented environment ever, Mac OS X is delivering more software innovation than our industry has seen in a decade.”



Dynamic and constantly-changing nature of Web content

- Google-search “World Trade Center” on September 11, 2001 returns descriptive pages about the building itself as top results.
 - Based on a model which periodically collected Web pages (days or weeks earlier), and indexed them
- Main search engines built specialized “[News Search](#)” features, which collect articles continuously from a number of news sources,
 - to answer queries about news stories minutes after they appear.
- Now news search features are partly integrated into the core parts of the search engine interface
- Emerging Web sites such as Twitter continue to fill in the spaces that exist between static content and real-time awareness.





Lauren

All Images Maps News Videos More Search tools

About 13,700,000 results (0.63 seconds)

Queen's College On The Web

www.qc.edu.hk/ ▾
Provides an introduction, a list of in-house organisations, and information on school events.
[Intranet](#) • [Administration](#) • [Untitled Document](#)

Queen's College, Hong Kong - Wikipedia, the free ...

https://en.wikipedia.org/wiki/Queen's_College,_Hong_Kong
Queen's College (皇仁書院), initially named The Government Central School (中央書院) in 1862, later renamed as Victoria College (皇后書院) in 1889, is a sixth ...
[Brief history](#) • [School song](#) • [School Motto](#) • [Enrollment and medium of ...](#)

Queen's College - Facebook

www.facebook.com/... > ... > [High School](#) ▾ [Translate this page](#)
Queen's College, Hong Kong, Hong Kong. 6525 likes · 320 talking about this · 10756 were here. One of the most prestigious schools over the territory...

Queens College

www.qc.cuny.edu/ ▾
A senior college of The City University of New York. Located in Queens.

首頁: QCOBA

www.qcoba.hk/ ▾
[行政及財務部](#) · [教育及發展部](#) · [會員事務部](#) · [學生事務部](#) · [文化及傳訊部](#) · [法律事務部](#) · [活動籌備委員會](#) · [其他人員](#) · [學校管理](#) · [皇仁書院](#) · [皇仁舊生會夜中學](#) · [皇仁舊生 ...](#)

1862 — Queen's College 150th Anniversary Open Days ...


https://www.youtube.com/watch?v=wCnv_qtp_o
Apr 6, 2012 · Uploaded by QCNouveau
Queen's College, the first government school in Hong Kong, was set up in 1862, and for a century and a half ...

Queen's College high scorers top university entrance exams ...

www.scmp.com/.../queens-college-high-scorers-top-university-entrance-... ▾
Jul 15, 2013 - Queen's College in Causeway Bay had the most top scorers, with two students getting seven 5**. Saint Paul's Co-Educational College and ...

Queens' College, University of Cambridge

www.queens.cam.ac.uk/ ▾



Map data ©2016 Google

Queen's College, Hong Kong ★

[Website](#) [Directions](#)

Government school in Hong Kong

Queen's College, initially named The Government Central School in 1862, later renamed as Victoria College in 1889, is a sixth form college for boys with a secondary school attached. [Wikipedia](#)

Address: 高士威道120號
Phone: 2576 1992
Founded: 1862
Color: Red

[Suggest an edit](#) · [Own this business?](#)

[Send to your phone](#) [Send](#)






Reviews

31 Google reviews

[Write a review](#) [Add a photo](#)

People also search for

[View 15+ more](#)

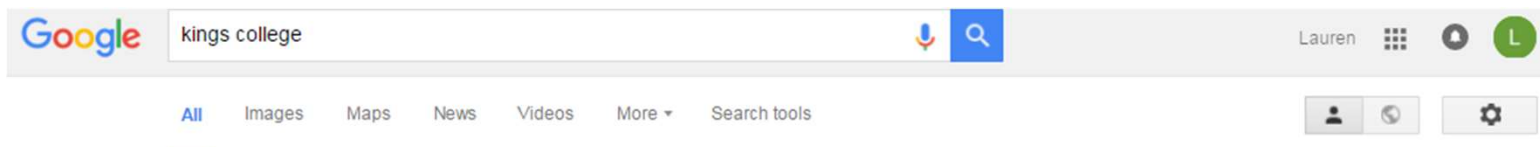


[Wah Yan](#) [Wah Yan](#) [Belilios](#) [St. Paul's](#) [Queen's](#)

2019/20 Term 2

CSCI4190 by Laiwan Chan

11



About 27,100,000 results (0.73 seconds)

King's College London - Home

www.kcl.ac.uk/

King's is one of the world's leading research and teaching universities based in the heart of London.

[Undergraduate study](#) • [Internal](#)

King's College School Official Website

www.kings.edu.hk/ • [Translate this page](#)

一所官立中學,幫助學生充分發現和發展自己的潛能,提供一個在德、智、體、美均衡發展的教育學習環境。

[King's College School Official ...](#) • [Gallery](#) • [Contact Us](#) • [School Song](#)

King's College, Cambridge - University of Cambridge

www.kings.cam.ac.uk/

Official site. Contains information about education, research, the facilities, the chapel, and the choir.

King's College London - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/King's_College_London

King's College London is a public research university located in London, United Kingdom, and a constituent college of the federal University of London.

King's College, Hong Kong - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/King's_College,_Hong_Kong

King's College, Hong Kong (Chinese: 英皇書院), often informally referred to simply as King's, is a single-sex boys' secondary school located at 63A Bonham ...

School Magazine: The Fig Tree

Number of students: ~1,200 students

Principal: Mrs. Chan Woo Mei-hou, Nan...

Staff: 69

In the news

Glendora, FedEx partner on King's College history book

[Vanguard](#) - 3 hours ago

GLENDORA International Limited and FedEx Red Star Express have entered into a partnership with the King's College Old Boys' Association ...

Critically ill man taken to Kings College Hospital, London, after car crash in Wellesley Road, Ashford

[Kent Online](#) - 2 days ago



King's College London ★

Public university in London, England

[Website](#)

[Directions](#)

King's College London is a public research university located in London, United Kingdom, and a constituent college of the federal University of London. [Wikipedia](#)

Address: Strand, London WC2R 2LS, United Kingdom

Acceptance rate: 13% (2014)

Phone: +44 20 7836 5454

Founded: 1829

Awards: The Queen's Award for Enterprise, International Trade

Founders: George IV of the United Kingdom, Arthur Wellesley, 1st Duke of Wellington

[Suggest an edit](#)

[Send to your phone](#)

[Send](#)

Profiles



[Facebook](#)

[Notable alumni](#)

[View 45+ more](#)



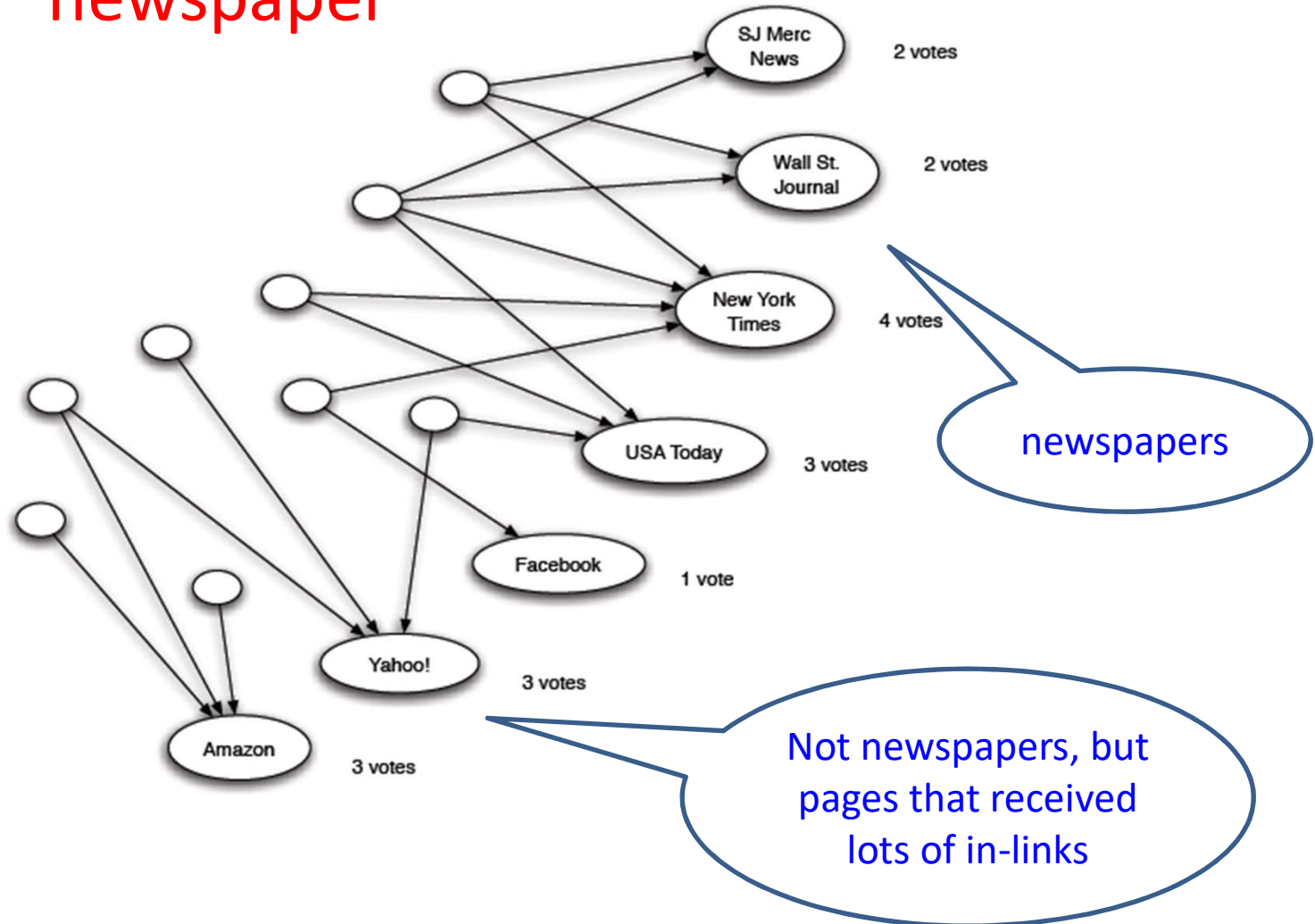
Voting by In-Links

- Vote by counting in-links
 - If a page receives many links from other relevant pages, then it is receiving a kind of collective **endorsement**.
- Each individual link may have many possible meanings
 - off-topic
 - criticism rather than endorsement
 - a paid advertisement



Example

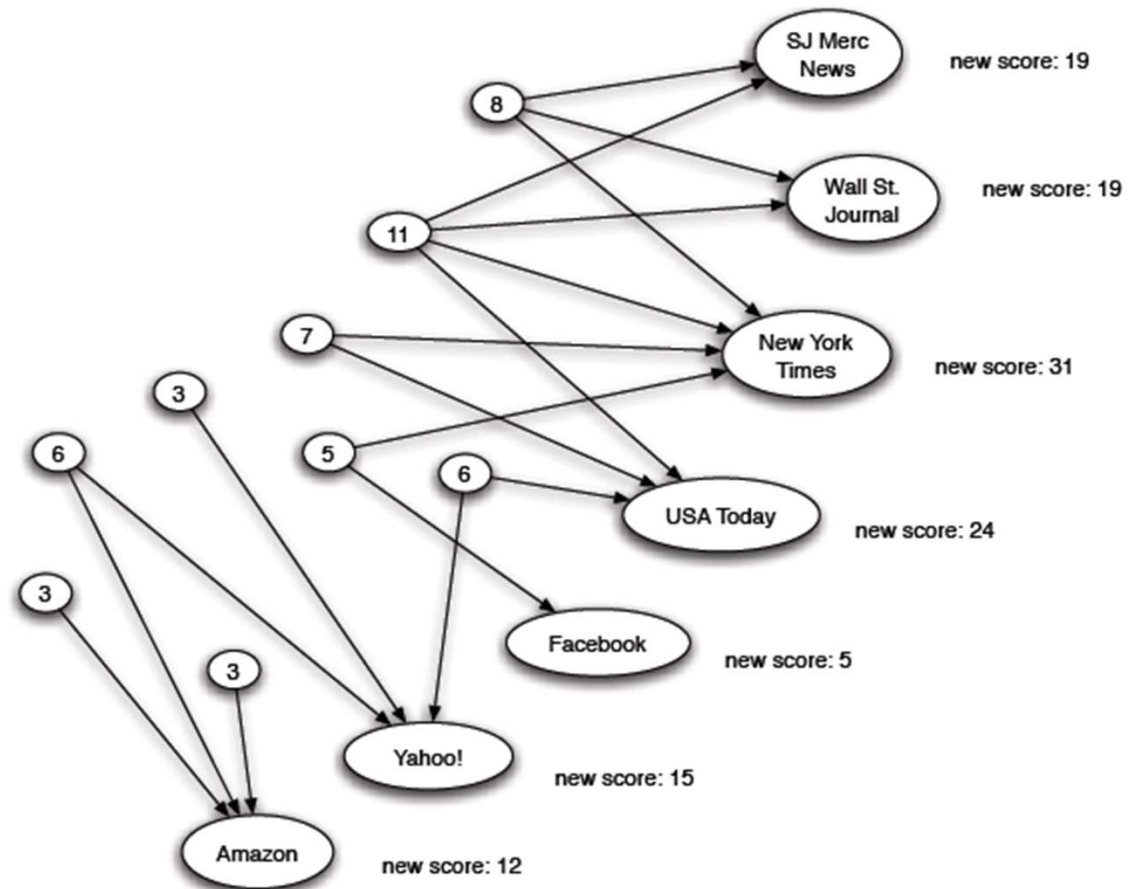
Query : "newspaper"





A List-Finding Technique

Give each page's vote
a weight equal to its
value as a list





The Principle of Repeated Improvement

- Why stop here?
- We can re-weight the votes against with more refined estimates
- The process can go back and forth forever



Hubs and Authorities

(also known as Hyperlink-induced
Topic Search, HITS)

- *Authorities* are pages that are recognized as providing significant, trustworthy, and useful information on a topic.
- *Hubs* are index pages that provide lots of useful links to relevant content pages (topic authorities).



hubs

Degree-awarding higher
education institutions

www.edb.gov.hk/

Aspirations for the Higher
Education System in Hong Kong

www.ugc.edu.hk/

Education - Overview | Census
and Statistics Department

www.censtatd.gov.hk/

Hong Kong Fact Sheets -
Education - Gov

www.gov.hk/

authorities

CUHK

HKU

HKUST

:



Hubs and Authorities

- For each page p ,
 - $auth(p)$: its value as a potential authority
 - $hub(p)$: its value as a potential hub
- **Authority Update Rule:**
 - For each page p , update $auth(p)$ to be the **sum of the hub scores** of all pages that point to it.

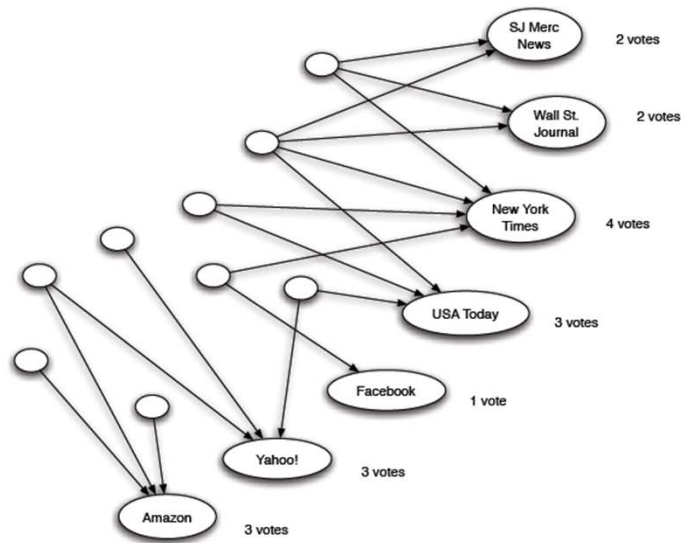
$$auth(p) = \sum_{(i,p) \in Edges} hub(i)$$

- **Hub Update Rule:**
 - For each page p , update $hub(p)$ to be the **sum of the authority scores** of all pages that it points to.

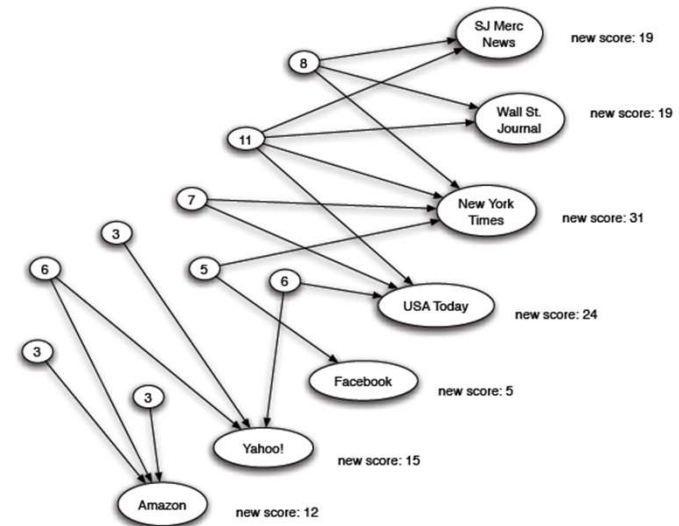
$$hub(p) = \sum_{(p,i) \in Edges} auth(i)$$



- Start with all hub scores and all authority scores equal to 1
- Choose k , the number of steps.
- Perform a sequence of k hub-authority updates
 - First apply the Authority Update Rule to the current set of scores
 - Then apply the Hub Update Rule to the resulting set of scores
- Normalize the scores to make them small
 - divide each authority score by the sum of all authority scores
 - divide each hub score by the sum of all hub scores



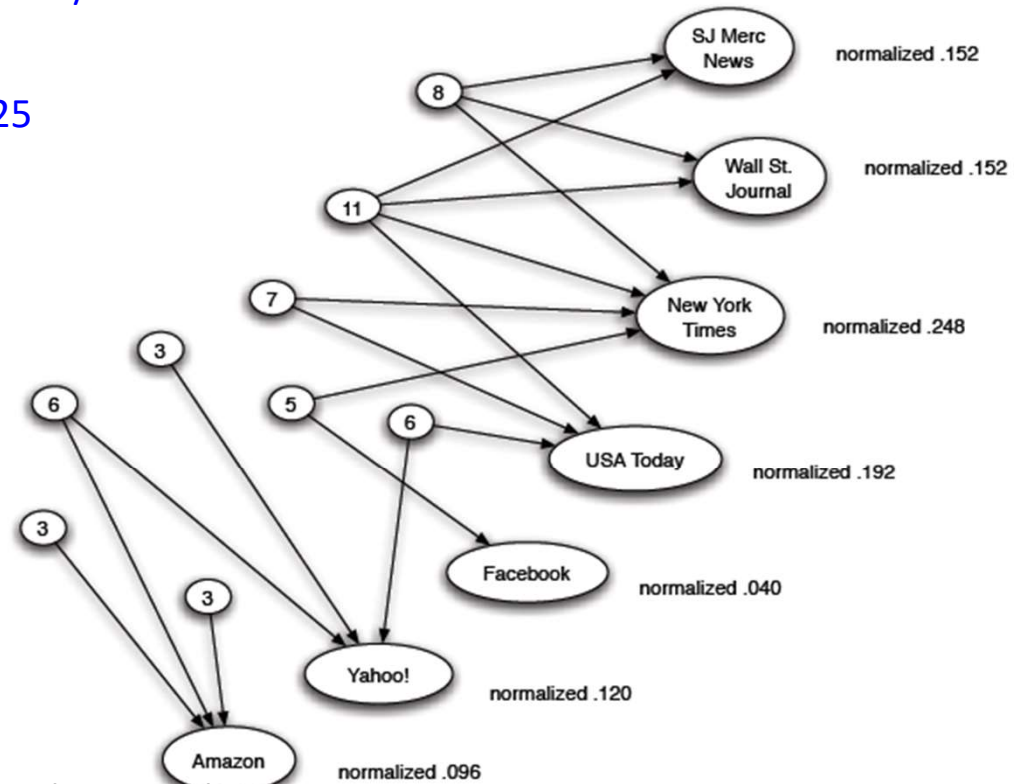
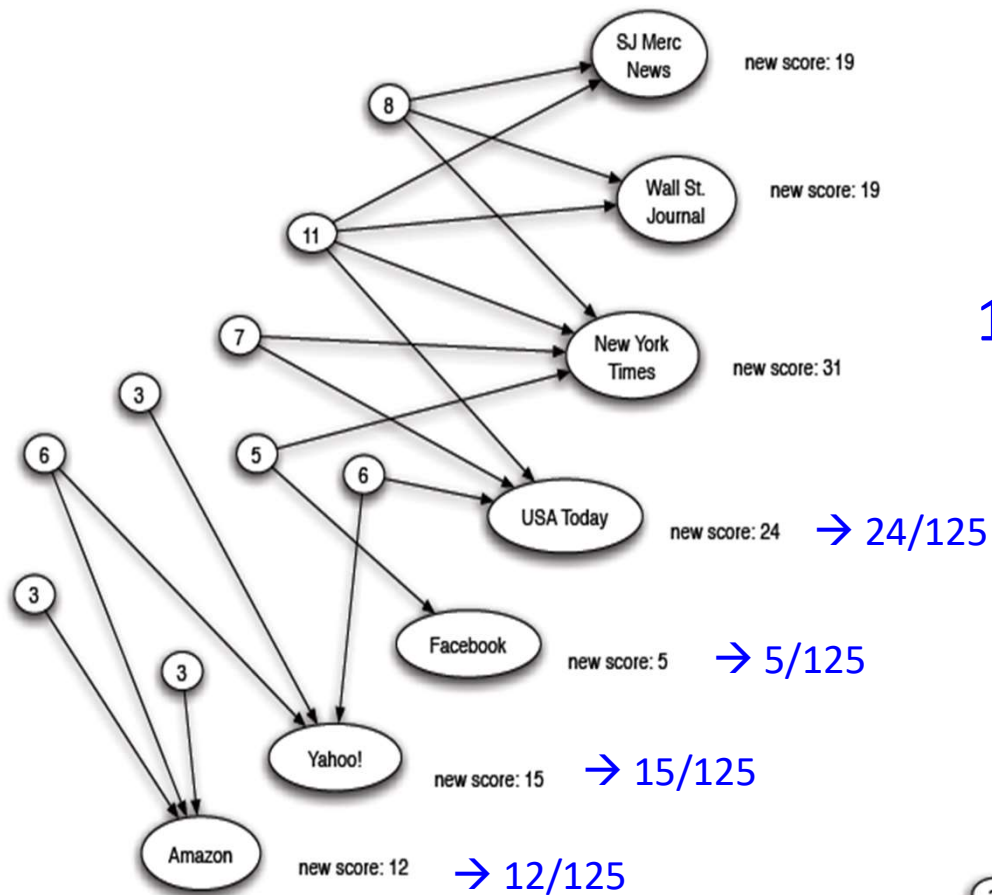
Authority Update Rule



Hub Update Rule

Normalization

$$19+19+31+24+5+15+12=125$$

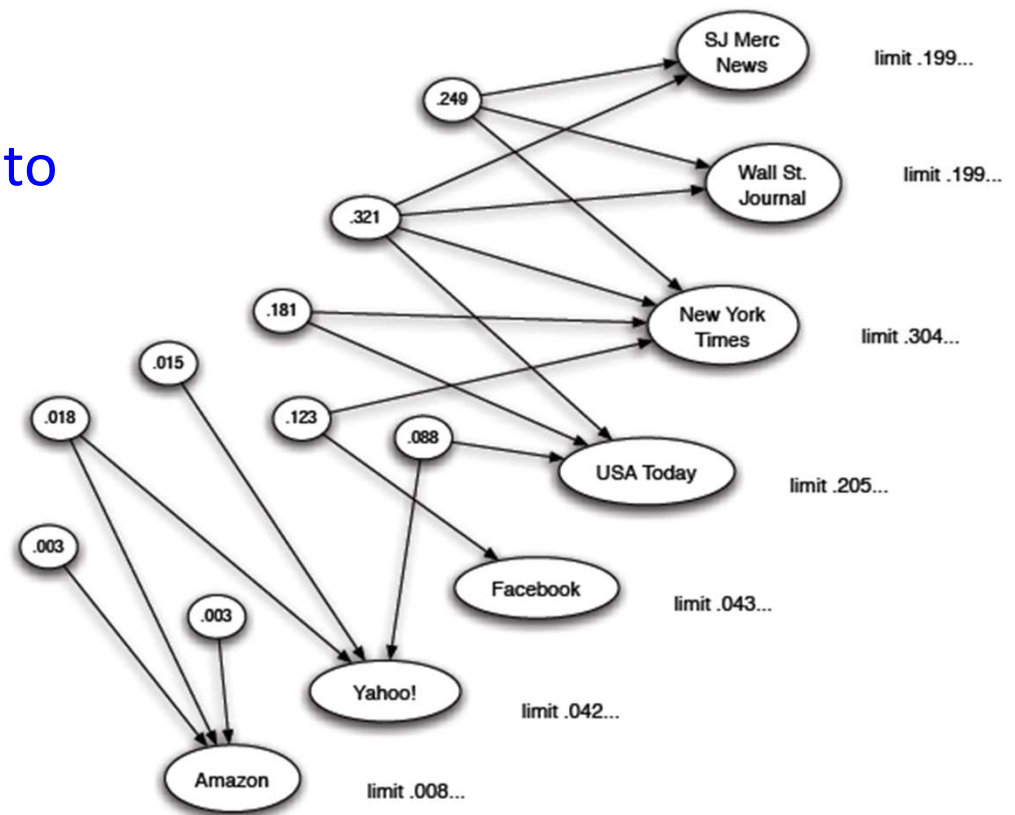




When k is large

Normalized values actually converge to limits as k goes to infinity.

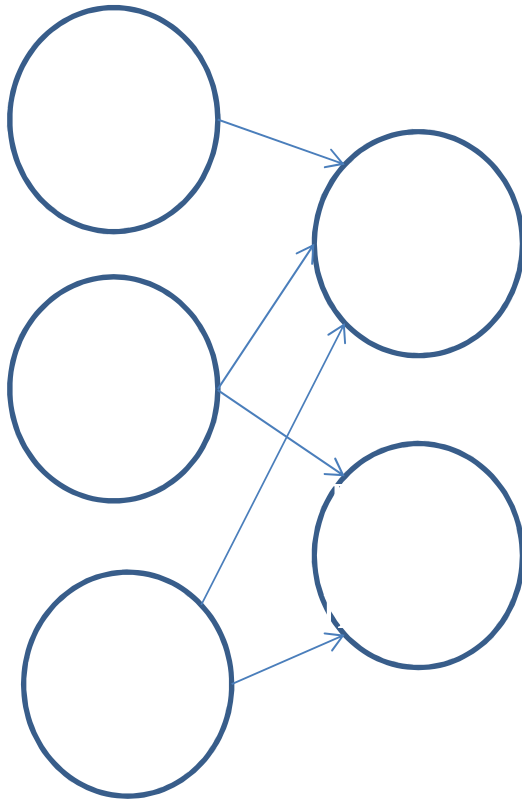
The same limiting values no matter what we choose as the initial hub and authority values





- M = adjacent matrix
- $auth(p) = \sum_{(i,p) \in Edges} hub(i)$
 $\Leftrightarrow auth = M^T hub$
- $hub(p) = \sum_{(p,i) \in Edges} auth(i)$
 $\Leftrightarrow hub = M auth$

hub and auth will converge to the corresponding principle eigenvectors



$$M = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad M^*M^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

The eigenvalues are $\lambda = \frac{5 \pm \sqrt{17}}{2} = 4.5616, 0.4384$

The eigenvectors are $\begin{bmatrix} (\lambda - 4)c \\ c \\ c \end{bmatrix}$

The eigenvectors are $\begin{bmatrix} 0.3690 \\ 0.6572 \\ 0.6572 \end{bmatrix}$ and $\begin{bmatrix} -0.9294 \\ 0.2610 \\ 0.2610 \end{bmatrix}$

Principle Eigenvector(M^*M^T) = $\begin{bmatrix} 0.3690 \\ 0.6572 \\ 0.6572 \end{bmatrix}$

Re-scaling the principle eigenvector = $\begin{bmatrix} 0.2192 \\ 0.3904 \\ 0.3904 \end{bmatrix}$



PageRank

- For queries with a commercial aspect
 - Competing firms will not link to each other
 - They are not directly endorsing each other
- On the Web
 - Endorsement is best viewed as passing directly from one prominent page to another
 - A page is important if it is cited by other important pages.

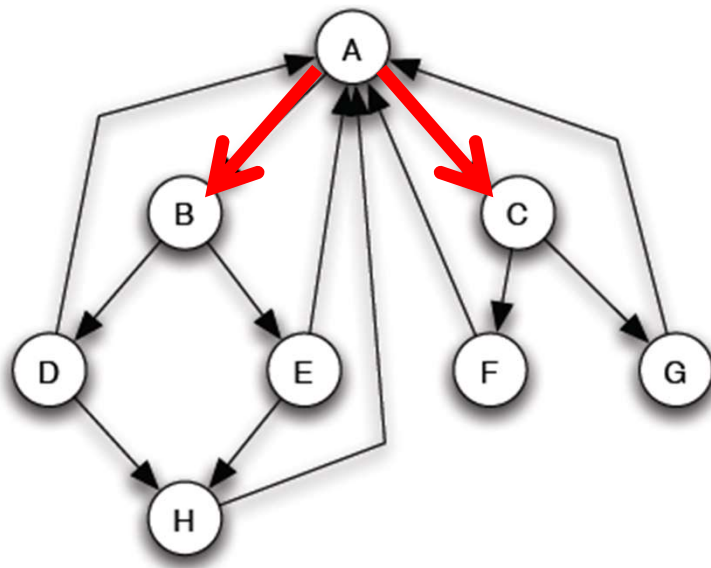


The basic definition of PageRank

- In a network with n nodes, all nodes have the same initial PageRank = $1/n$.
- Choose k , the number of steps.
- Perform a sequence of k updates to the PageRank values
 - Basic PageRank Update Rule
 - Each page divides its current PageRank equally across its out-going links, and passes these equal shares to the pages it points to.
 - If a page has no out-going links, it passes all its current PageRank to itself.
 - Each page updates its new PageRank to be the sum of the shares it receives.



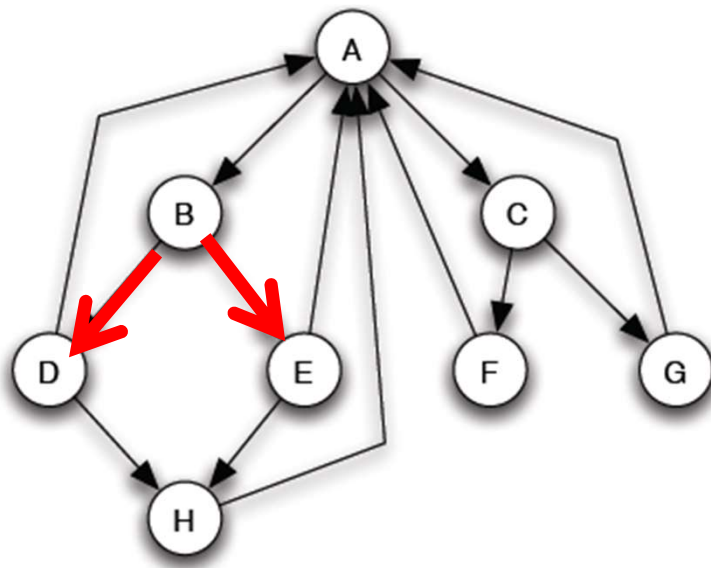
Step	A	B	C	D	E	F	G	H
0	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8



	A	B	C	D	E	F	G	H
A		1/16	1/16					
B				1/16	1/16			
C						1/16	1/16	
D	1/16							1/16
E	1/16							1/16
F	1/8							
G	1/8							
H	1/8							



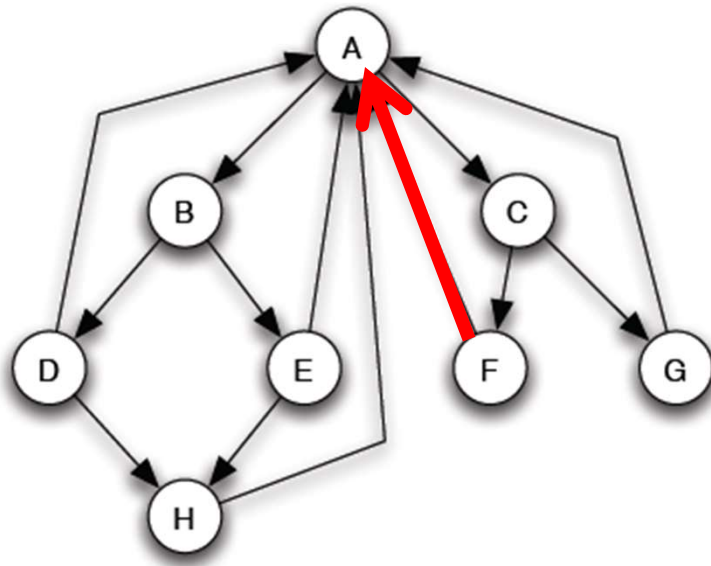
Step	A	B	C	D	E	F	G	H
0	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8



	A	B	C	D	E	F	G	H
A		1/16	1/16					
B				1/16	1/16			
C						1/16	1/16	
D	1/16							1/16
E	1/16							1/16
F	1/8							
G	1/8							
H	1/8							



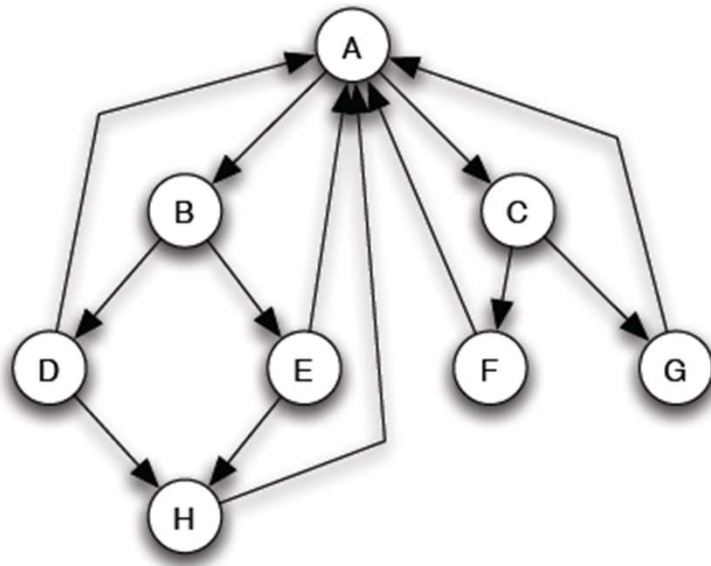
Step	A	B	C	D	E	F	G	H
0	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8



	A	B	C	D	E	F	G	H
A		1/16	1/16					
B				1/16	1/16			
C						1/16	1/16	
D	1/16							1/16
E	1/16							1/16
F	1/8							
G	1/8							
H	1/8							



Step	A	B	C	D	E	F	G	H
0	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

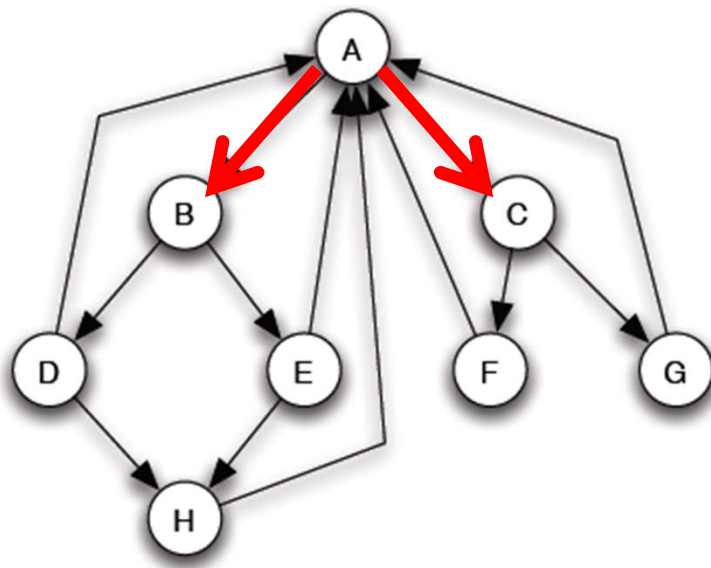


	A	B	C	D	E	F	G	H
A		1/16	1/16					
B				1/16	1/16			
C						1/16	1/16	
D	1/16							1/16
E	1/16							1/16
F	1/8							
G	1/8							
H	1/8							

total	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8



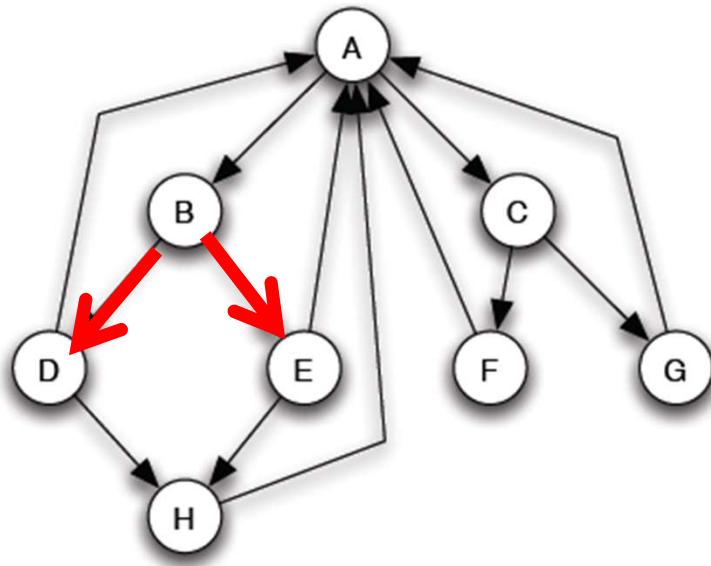
Step	A	B	C	D	E	F	G	H
1	$\frac{1}{2}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$



	A	B	C	D	E	F	G	H
A		$\frac{1}{4}$	$\frac{1}{4}$					
B				$\frac{1}{32}$	$\frac{1}{32}$			
C						$\frac{1}{32}$	$\frac{1}{32}$	
D	$\frac{1}{32}$							$\frac{1}{32}$
E	$\frac{1}{32}$							$\frac{1}{32}$
F	$\frac{1}{16}$							
G	$\frac{1}{16}$							
H	$\frac{1}{8}$							



Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8

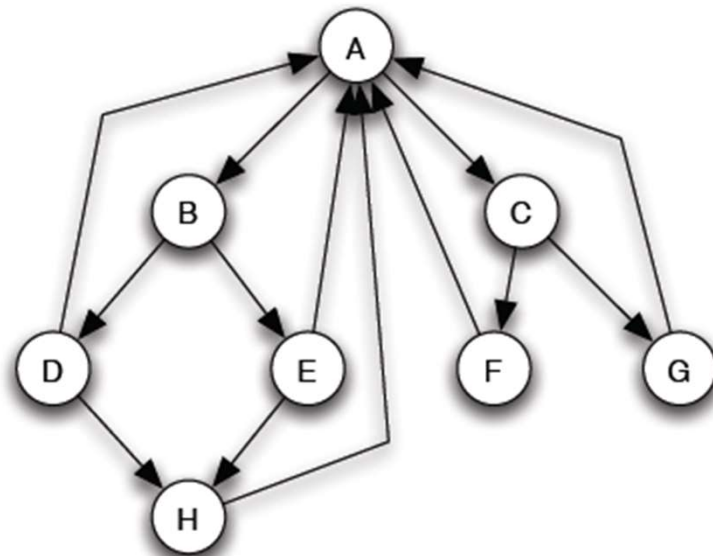


	A	B	C	D	E	F	G	H
A		$\frac{1}{4}$	$\frac{1}{4}$					
B				1/32	1/32			
C						1/32	1/32	
D	1/32							1/32
E	1/32							1/32
F	1/16							
G	1/16							
H	1/8							

total	A	B	C	D	E	F	G	H
1	5/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16



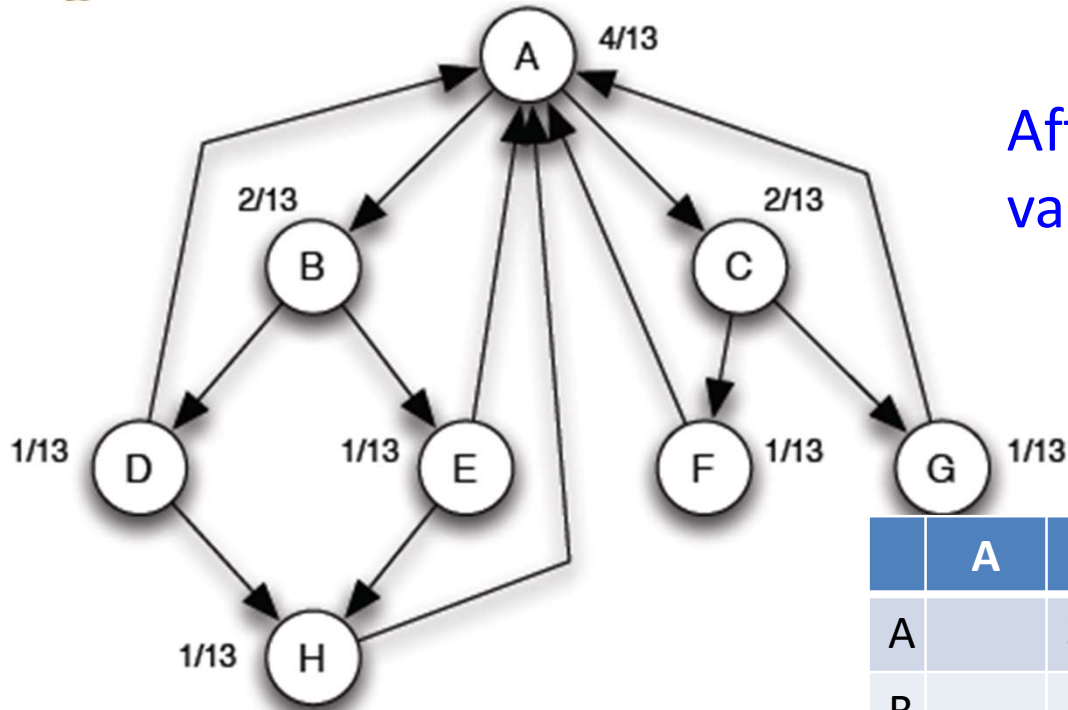
Step	A	B	C	D	E	F	G	H
0	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$	$1/8$
1	$1/2$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/8$
2	$5/16$	$1/4$	$1/4$	$1/32$	$1/32$	$1/32$	$1/32$	$1/16$
...								
	$4/13$	$2/13$	$2/13$	$1/13$	$1/13$	$1/13$	$1/13$	$1/13$





Equilibrium PageRank values

After a number of steps, values converge to limits.



	A	B	C	D	E	F	G	H
A		2/13	2/13					
B				1/13	1/13			
C						1/13	1/13	
D	1/26							1/26
E	1/26							1/26
F	1/13							
G	1/13							
H	1/13							



- Initialize ranks $R_0(j)$ for all nodes j
- N_j = number of out links of node j
- While not converged

For each vertex i

$$R_{k+1}(i) = \sum_{(j,i) \in E} \frac{R_k(j)}{N_j}$$

N_j is not zero

End for

End while

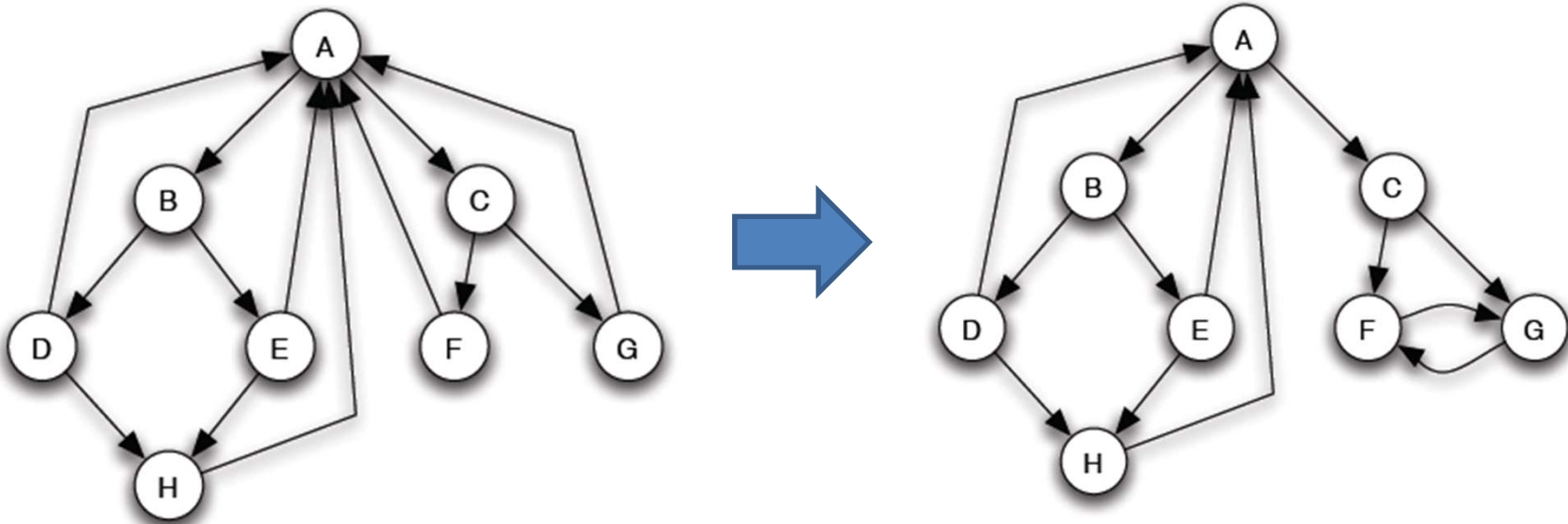


M = adjacent matrix

- HITS
 - $auth(p) = \sum_{(i,p) \in Edges} hub(i)$
 $\Leftrightarrow auth = M^T hub$
 - $hub(p) = \sum_{(p,i) \in Edges} auth(i)$
 $\Leftrightarrow hub = M auth$
- Pagerank
 - $r(p) = \sum_{(i,p) \in Edges} r(i)$
 $\Leftrightarrow r_{k+1} = M^T r_k$



Modify the network



Not the
result we
want

equilibrium
 $\frac{1}{2}$ for F
 $\frac{1}{2}$ for G
0 for others



Scaling Factor

- **Scaled PageRank Update Rule:**
 - First apply the Basic PageRank Update Rule.
 - Then scale down all PageRank values by a factor of s .
 - This means that the total PageRank in the network has shrunk from 1 to s .
 - We divide the residual $1 - s$ units of PageRank equally over all nodes, giving $(1 - s)/N$ to each.
- s usually chosen to be between 0.8 and 0.9.



- Initialize ranks $R_0(j)$ for all nodes j
- N_j = number of out links of node j
- While not converged

For each vertex i

$$R_{k+1}(i) = \frac{(1-s)}{N} + s \sum_{(j,i) \in E} \frac{R_k(j)}{N_j}$$

End for

End while

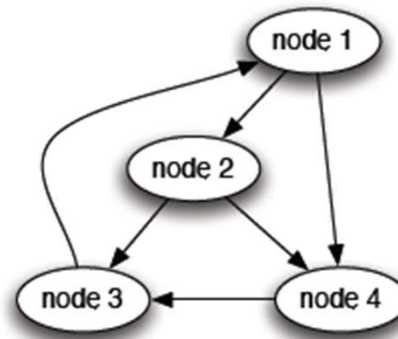


Node	1	2	3	4
1	0	0.5	0	0.5
2	0	0	0.5	0.5

$s = 0.8$

From node 1,
0.8 units are divided equally
between nodes 2 and 4
0.2 units are divided equally
over all nodes.

Node	1	2	3	4
1	0.05	0.45	0.05	0.45
2	0.05	0.05	0.45	0.45



$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$



$$R_{k+1}(i) = \frac{(1-s)}{N} + s \sum_{(j,i) \in E} \frac{R_k(j)}{N_j}$$

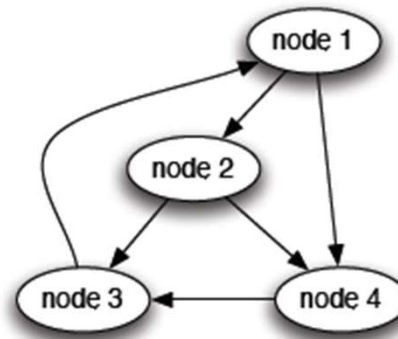


Node	1	2	3	4
1	0	0.5	0	0.5
2	0	0	0.5	0.5

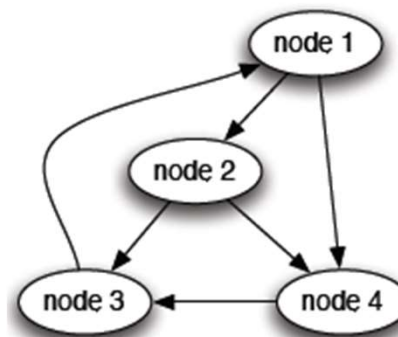
$s = 0.8$

From node 1,
0.8 units are divided equally
between nodes 2 and 4
0.2 units are divided equally
over all nodes.

Node	1	2	3	4
1	0.05	0.45	0.05	0.45
2	0.05	0.05	0.45	0.45
3	0.85	0.05	0.05	0.05
4	0.05	0.05	0.85	0.05



$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$



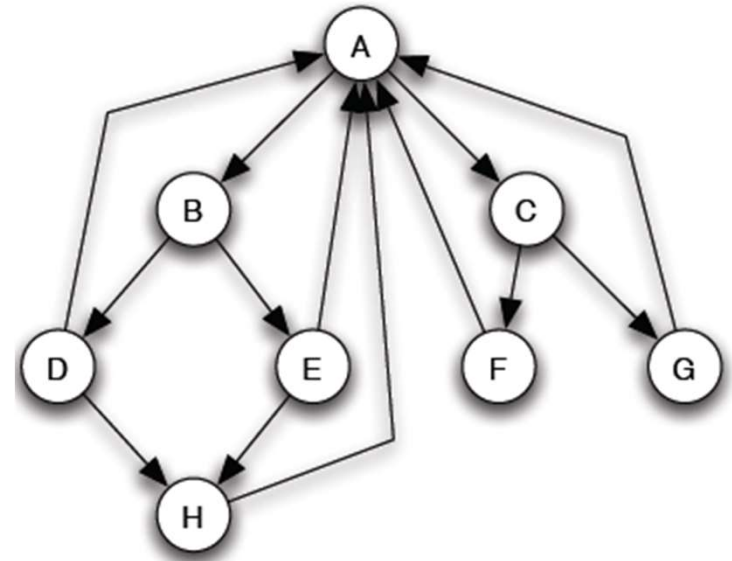
$$\begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$



Random walks

An equivalent formulation of PageRank

- Choose a page at random, and pick each page with equal probability.
- Follow links for a sequence of k steps: in each step, they pick a random outgoing link from their current page, and follow it to where it leads.





- **Claim** : The probability of being at a page X after k steps of this random walk is precisely the PageRank of X after k applications of the Basic PageRank Update Rule.



Combining links, text and usage data

- Links and text

- I am a student at [CUHK](#) ← anchor text that activate a hyperlink to a page about CUHK.
- Links with relevant anchor text could have heavier weightings when we process the hub/authority scores or PageRank Values

- Usage data

- The number of user clicks in previous search results is a good indication of its relevance.



Applications beyond the web

- Citation analysis
 - Impact factor = average number of citations received by a paper over the past two years
 - Citations from high-impact journals have heavier weightings → influence weights for journals.



- **Link analysis of US supreme court citations**
 - Citations are crucial in legal writings, to ground a decision in precedent and to explain the relation of a new decision to what has come before.

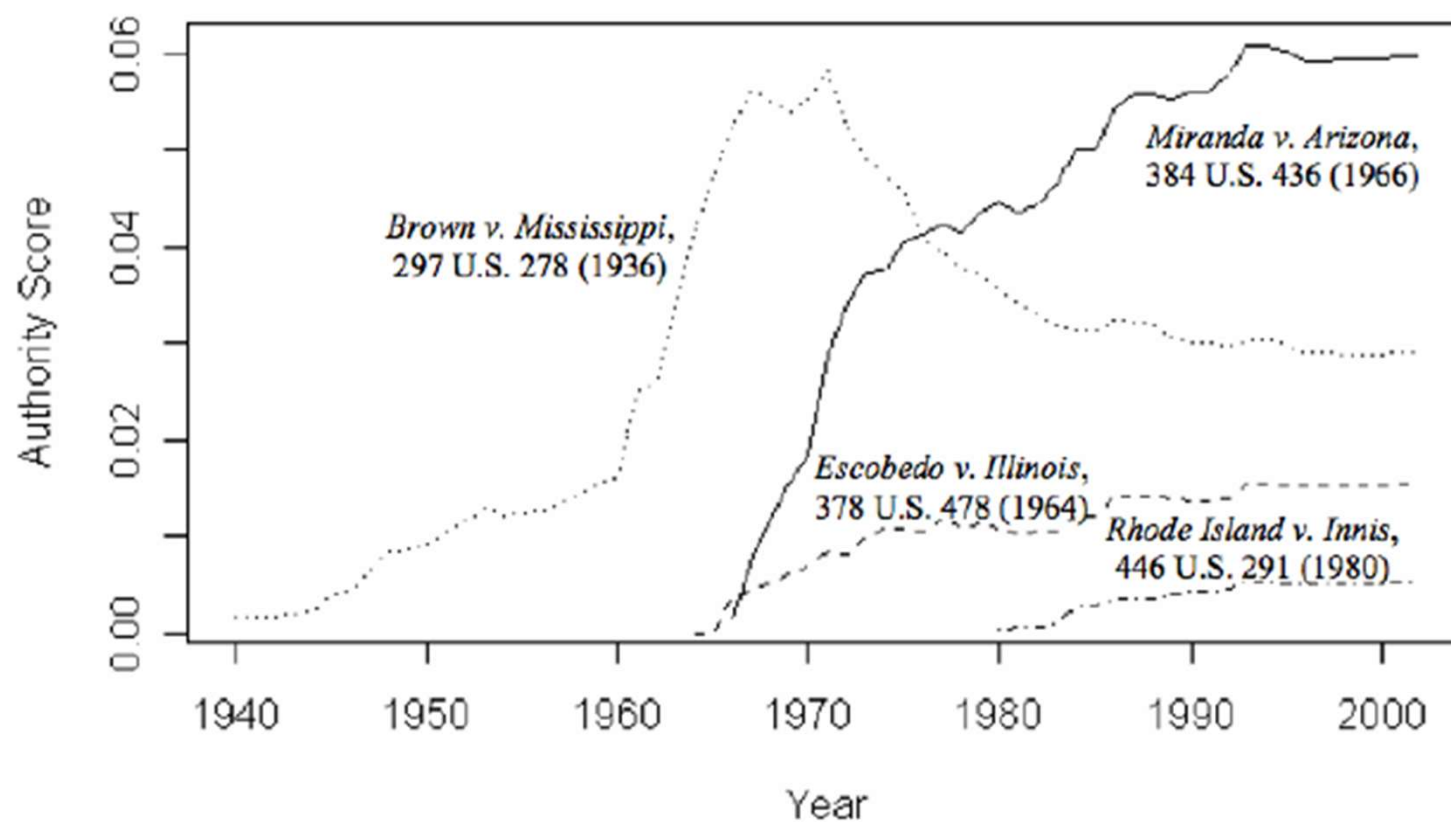


Figure 14.9: The rising and falling authority of key Fifth Amendment cases from the 20th century illustrates some of the relationships among them. (Image from [166].)

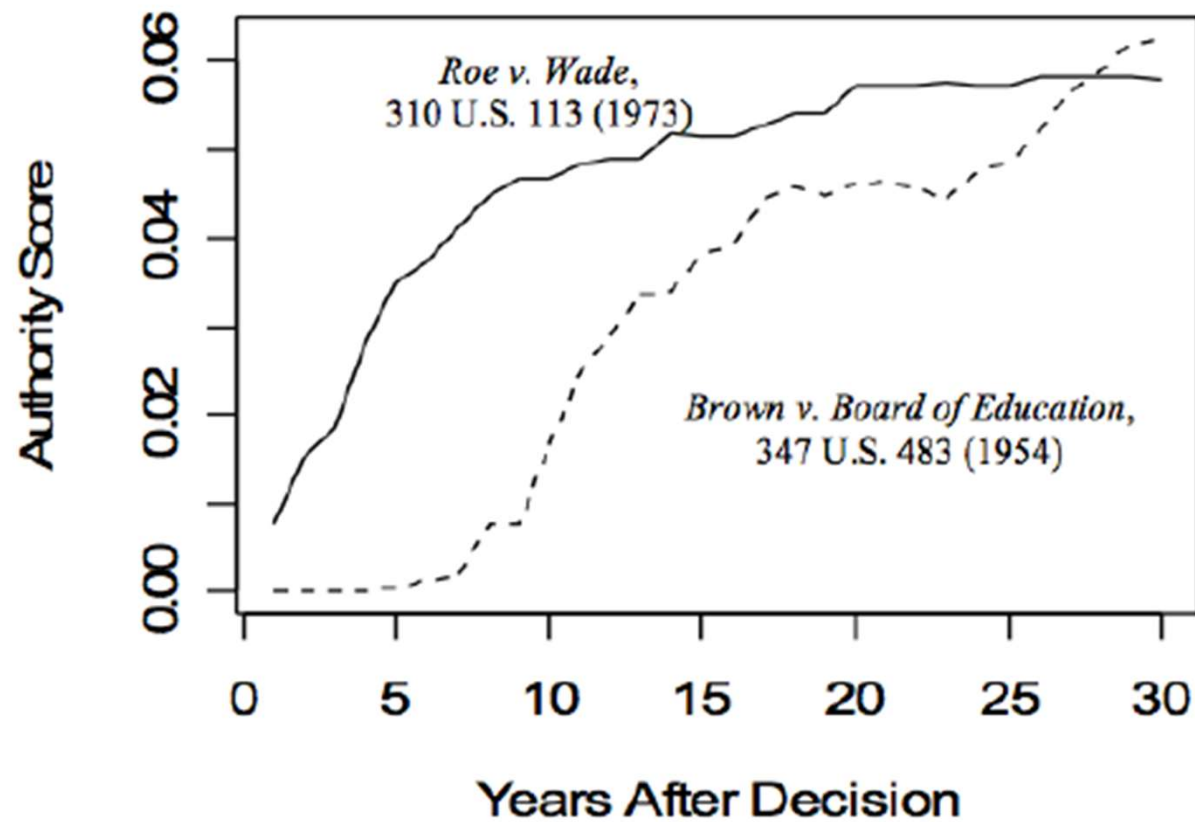


Figure 14.10: *Roe v. Wade* and *Brown v. Board of Education* acquired authority at very different speeds. (Image from [166].)