

Social and Information Networks

**CSCC46H, Fall 2020
Lecture 7**

Prof. Ashton Anderson
ashton@cs.toronto.edu

Logistics

A2 due today; A3 out tomorrow

Today

PageRank
Node centrality

How to Organize and Find Information?

Google university of toronto

All Maps Images News Videos More Settings Tools

About 404,000,000 results (0.87 seconds)

University of Toronto
<https://www.utoronto.ca/> ▼
U of T physicist Ursula Franklin staunchly opposed weapons of mass destruction. Today, her work remains as relevant as ever. **University of Toronto.**

Results from utoronto.ca


Programs
Programs of Study. Dive into your interests and develop your ...


International Students
Fees - Career Learning - Study Permits - ...


Future Students
That is what creates an exceptional learning ...

UToronto Webmail
[https://weblogin.utoronto.ca/...](https://weblogin.utoronto.ca/) Check your browser for a ...

Top stories from utoronto.ca


Toronto makes pitch to be world's 'Stem Cell City' with new, state-of-the-art lab
University of Toronto · 3 day...


Maclean's names U of T Canada's top school by reputation for 2018
University of Toronto · 4 day...


is world's top public university, fourth overall: National Taiwan University Ranking
University of Toronto · 5 day...

→ More for university of toronto

University of Toronto - Wikipedia
https://en.wikipedia.org/wiki/University_of_Toronto ▼
The **University of Toronto** is a public research university in Toronto, Ontario, Canada on the grounds that surround Queen's Park. It was founded by royal charter ...
[History](#) · [Academics](#) · [Research](#) · [Culture and student life](#)

University of Toronto World University Rankings | THE
<https://www.timeshighereducation.com/world-university-rankings/university-toronto> ▼
Find the latest world ranking position for **University of Toronto** and key information for prospective students here today.

How do you do this, exactly?

How to Organize Information?

How do you organize something as vast and messy as the Web?

First try: Human-curation

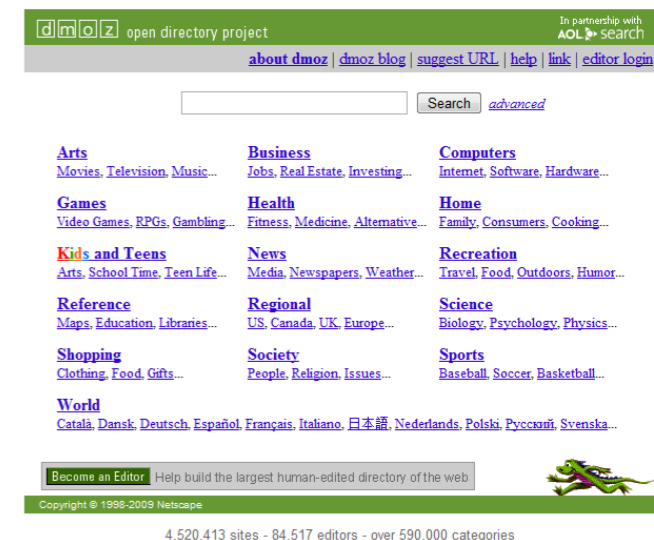
Web directories

Yahoo, DMOZ, LookSmart

How do you organize big collections of documents containing information?

Goes way back before the Web

Patents, Legal cases, Medical research



How to Organize Information?

It's a hard problem!

Given a relatively tiny keyword string, find ~5 most relevant and important documents out of 100K, 1M, 10M..., 10B... documents



How to Organize the Web?

How do you organize the Web?

First try: Human curation

Web directories

Yahoo, DMOZ, LookSmart

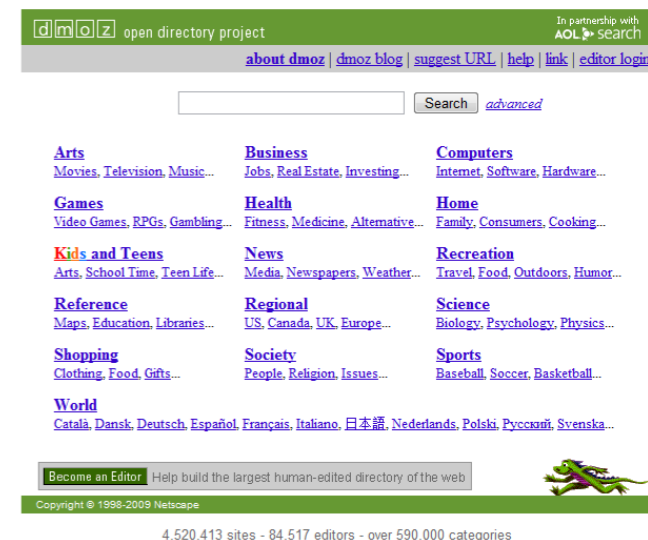
Second try: Web Search

Information Retrieval attempts to find relevant docs in a small and trusted set

Newspaper articles, Patents, etc.

But: The Web is huge, full of untrusted documents, random things, web spam, etc.

So we need a good way to rank webpages!

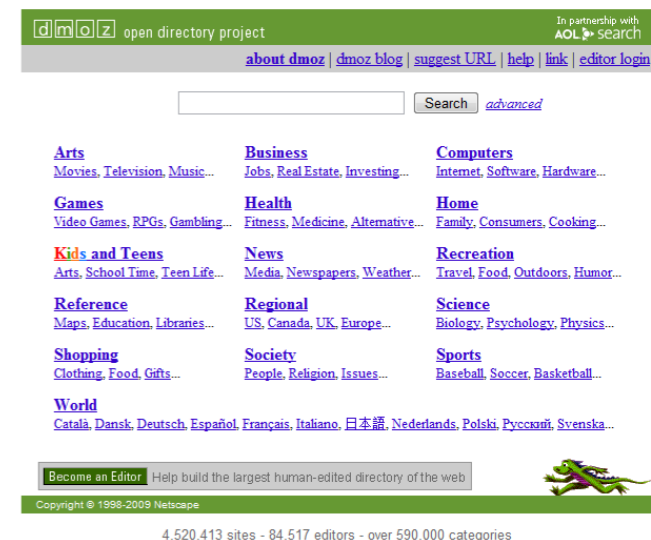


How to Organize the Web?

Search engines efficiently rank any query using only info from the Web

So the answer has to be intrinsically there somewhere...

How do we do this?



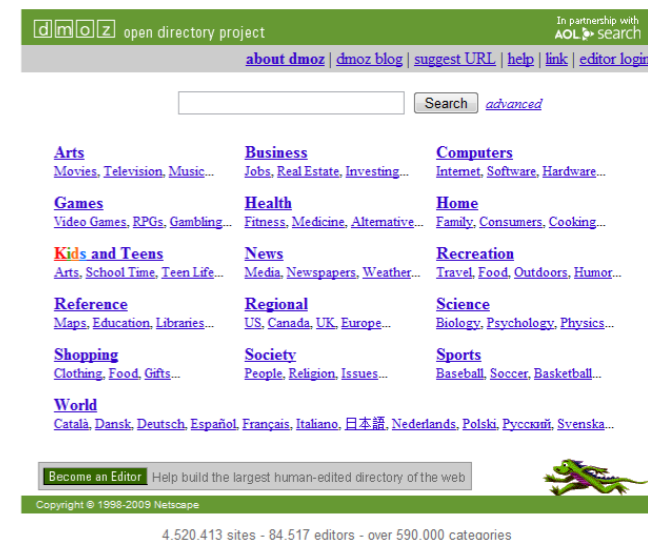
How to Organize the Web?

Information retrieval pre-Web problem: **scarcity**

Combing through thousands of documents to try to find the right ones

Search on the web problem: **abundance**

So many “relevant” pages with good text matches and seemingly high production values
What to trust?



A First Try

Use the content on the page!

Can you separate good, relevant pages from irrelevant or even malicious pages?

The screenshot shows a Google search for "university of toronto". The search bar is at the top with the Google logo on the left and a search icon on the right. Below the search bar are navigation links for "All", "Maps", "Images", "News", "Videos", "More", "Settings", and "Tools". The search results indicate "About 404,000,000 results (0.87 seconds)".

The first result is "University of Toronto" with the URL <https://www.utoronto.ca/>. Below this is a snippet: "U of T physicist Ursula Franklin staunchly opposed weapons of mass destruction. Today, her work remains as relevant as ever. University of Toronto." Below the snippet is a search bar for "Results from utoronto.ca".

There are four quick links:

- Programs**: Programs of Study. Dive into your interests and develop your ...
- International Students**: Fees - Career Learning - Study Permits - ...
- Future Students**: That is what creates an exceptional learning ...
- UToronto Webmail**: <https://weblogin.utoronto.ca/>... Check your browser for a ...

Below these is a section titled "Top stories from utoronto.ca" with three story cards:

- Toronto makes pitch to be world's 'Stem Cell City' with new, state-of-the-art lab**. University of Toronto - 3 day...
- Maclean's names U of T Canada's top school by reputation for 2018**. University of Toronto - 4 day...
- is world's top public university, fourth overall: National Taiwan University Ranking**. University of Toronto - 5 day...

Below the stories is a link: [More for university of toronto](#)

Below that is a result for "University of Toronto - Wikipedia" with the URL https://en.wikipedia.org/wiki/University_of_Toronto. The snippet reads: "The University of Toronto is a public research university in Toronto, Ontario, Canada on the grounds that surround Queen's Park. It was founded by royal charter ...". Below the snippet are links for "History", "Academics", "Research", and "Culture and student life".

At the bottom is a result for "University of Toronto World University Rankings | THE" with the URL <https://www.timeshighereducation.com/world-university-rankings/university-toronto>. The snippet reads: "Find the latest world ranking position for University of Toronto and key information for prospective students here today."

Key Idea

Nothing on the “right” page makes it stand out from the thousands of others

But it will very often be **linked to** by others!

The screenshot shows a Google search for "university of toronto". The search bar at the top contains the text "university of toronto" and the Google logo. Below the search bar, there are navigation tabs for "All", "Maps", "Images", "News", "Videos", "More", "Settings", and "Tools". The search results indicate "About 404,000,000 results (0.87 seconds)".

The first result is "University of Toronto" with the URL <https://www.utoronto.ca/>. A snippet below the URL reads: "U of T physicist Ursula Franklin staunchly opposed weapons of mass destruction. Today, her work remains as relevant as ever. University of Toronto." Below this is a search box for "Results from utoronto.ca".

There are four quick links: "Programs" (Programs of Study, Dive into your interests and develop your...), "International Students" (Fees - Career Learning - Study Permits - ...), "Future Students" (That is what creates an exceptional learning...), and "UToronto Webmail" (<https://weblogin.utoronto.ca/>... Check your browser for a...).

Below these are "Top stories from utoronto.ca" with three story cards:

- Toronto makes pitch to be world's 'Stem Cell City' with new, state-of-the-art lab** (University of Toronto - 3 day...)
- Maclean's names U of T Canada's top school by reputation for 2018** (University of Toronto - 4 day...)
- is world's top public university, fourth overall: National Taiwan University Ranking** (University of Toronto - 5 day...)

A link "More for university of toronto" is provided below the stories.

Further down, there is a "University of Toronto - Wikipedia" result with the URL https://en.wikipedia.org/wiki/University_of_Toronto and a snippet: "The University of Toronto is a public research university in Toronto, Ontario, Canada on the grounds that surround Queen's Park. It was founded by royal charter ...".

At the bottom, there is a "University of Toronto World University Rankings | THE" result with the URL <https://www.timeshighereducation.com/world-university-rankings/university-toronto> and a snippet: "Find the latest world ranking position for University of Toronto and key information for prospective students here today."

Web Search: 2 Challenges

2 challenges of web search:

(1) Web contains many sources of information

Who to “trust”?

Insight: Trustworthy pages may point to each other!

(2) What is the “best” answer to query “newspaper”?

No single right answer

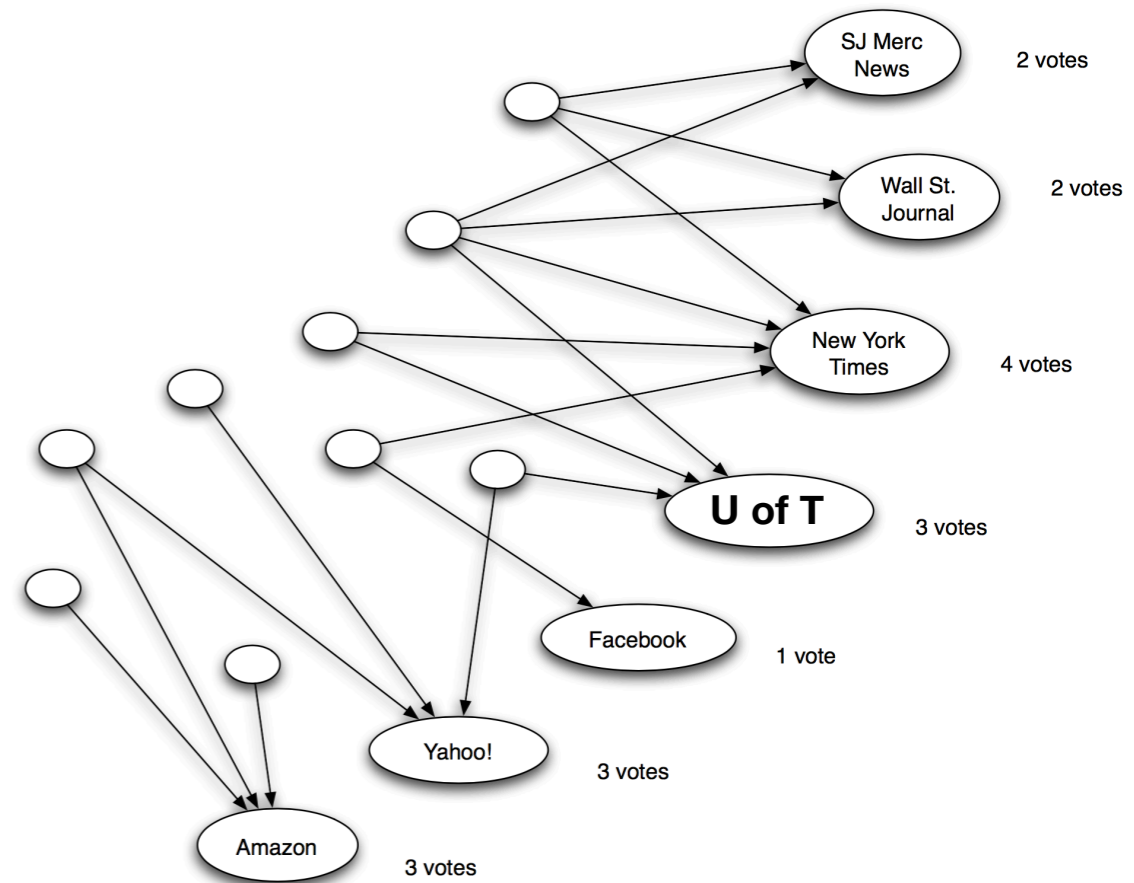
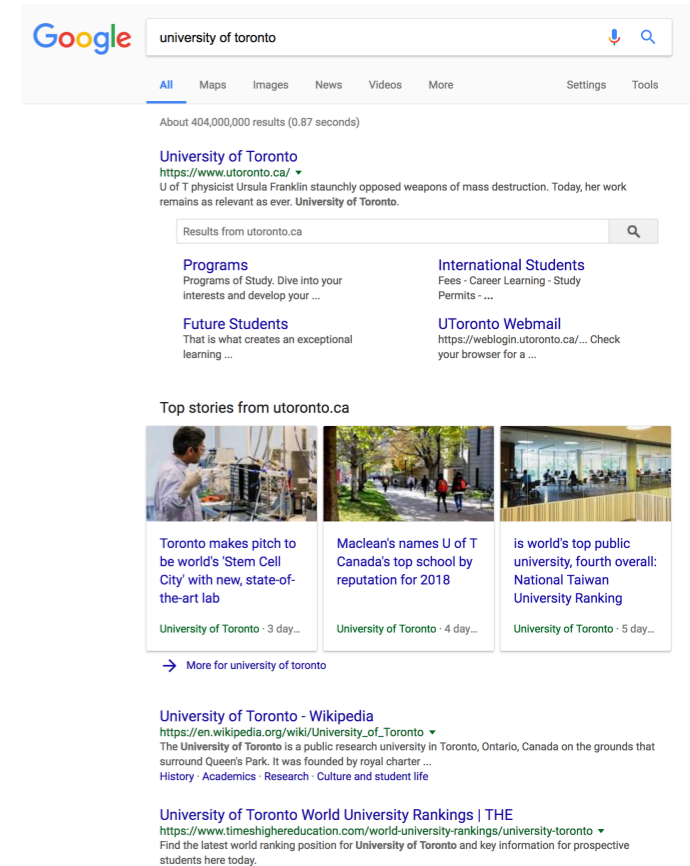
Insight: Pages that actually know about newspapers might all be pointing to many newspapers

Key Idea

Nothing on the “right” page makes it stand out from the thousands of others

But it will very often be **linked to** by others!

What’s a natural first thing to try if we want to harness the link structure?

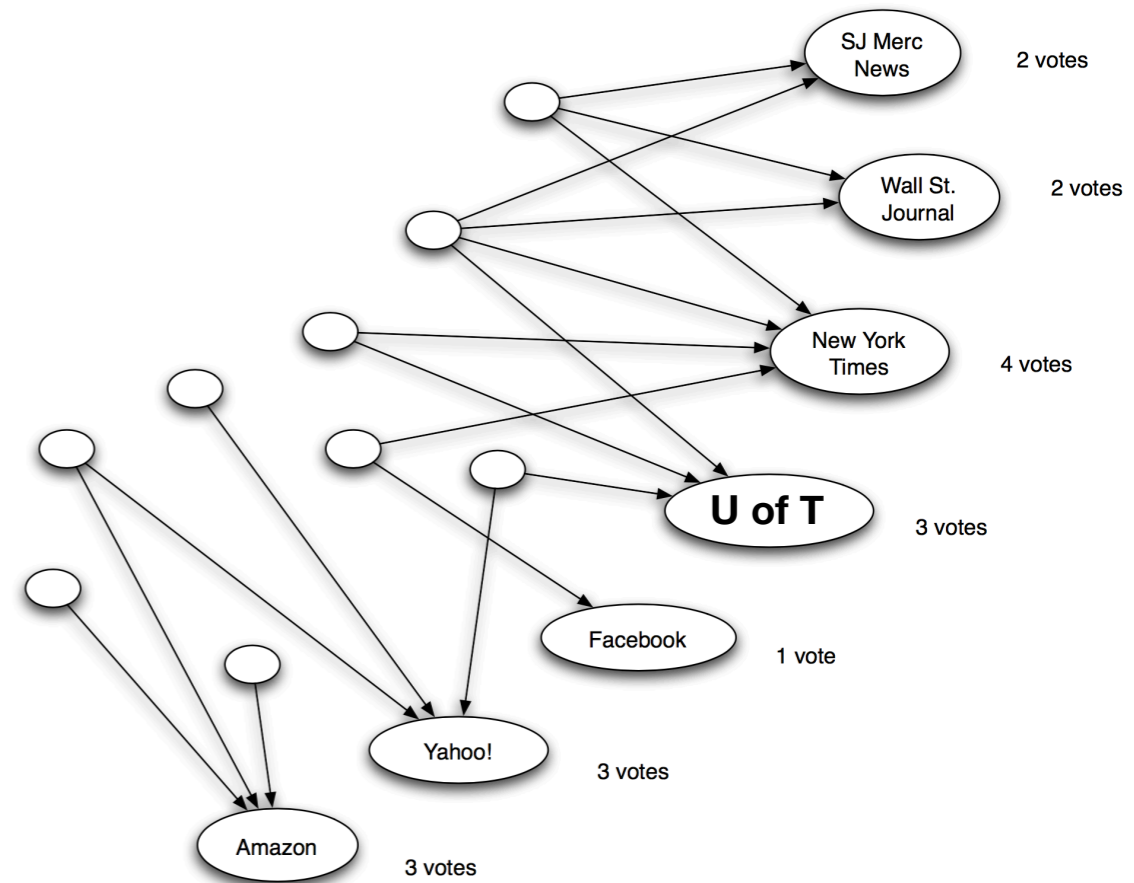
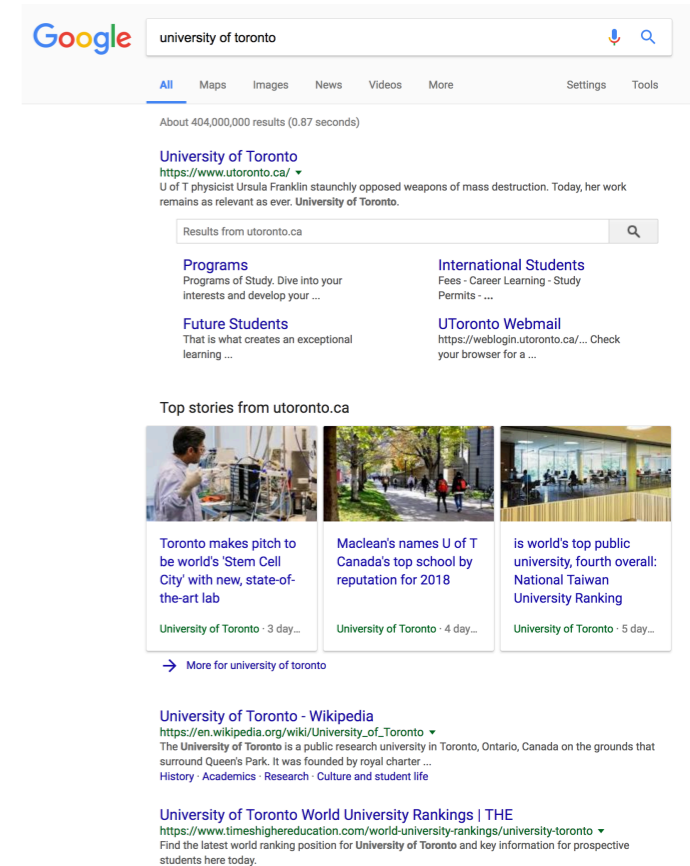


Key Idea

Nothing on the “right” page makes it stand out from the thousands of others

But it will very often be **linked to** by others!

Restrict to a relevant set and **count the in-links**



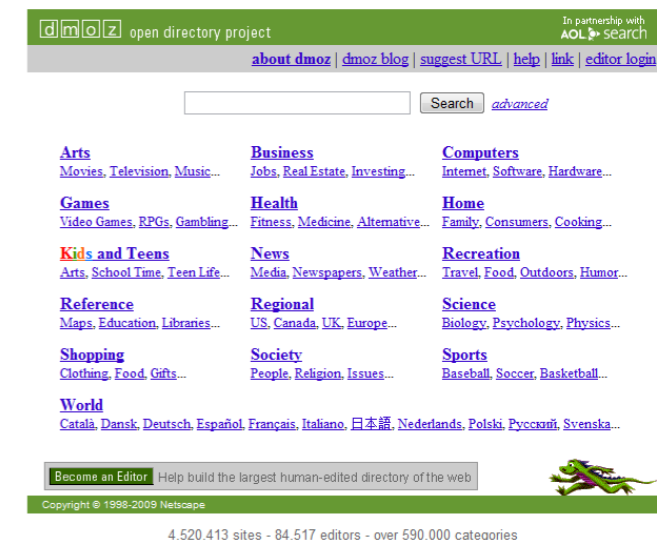
Idea: links as votes!

If I link to you, that's usually a good thing

1. Model the Web as a directed graph

2. Use the link structure to compute **importance values** of webpages

3. Use these importance values for **ranking**



Link Analysis Algorithms

We will cover the following link analysis algorithms to compute the importance or centrality of nodes in a graph:

Hubs and Authorities (HITS)

PageRank

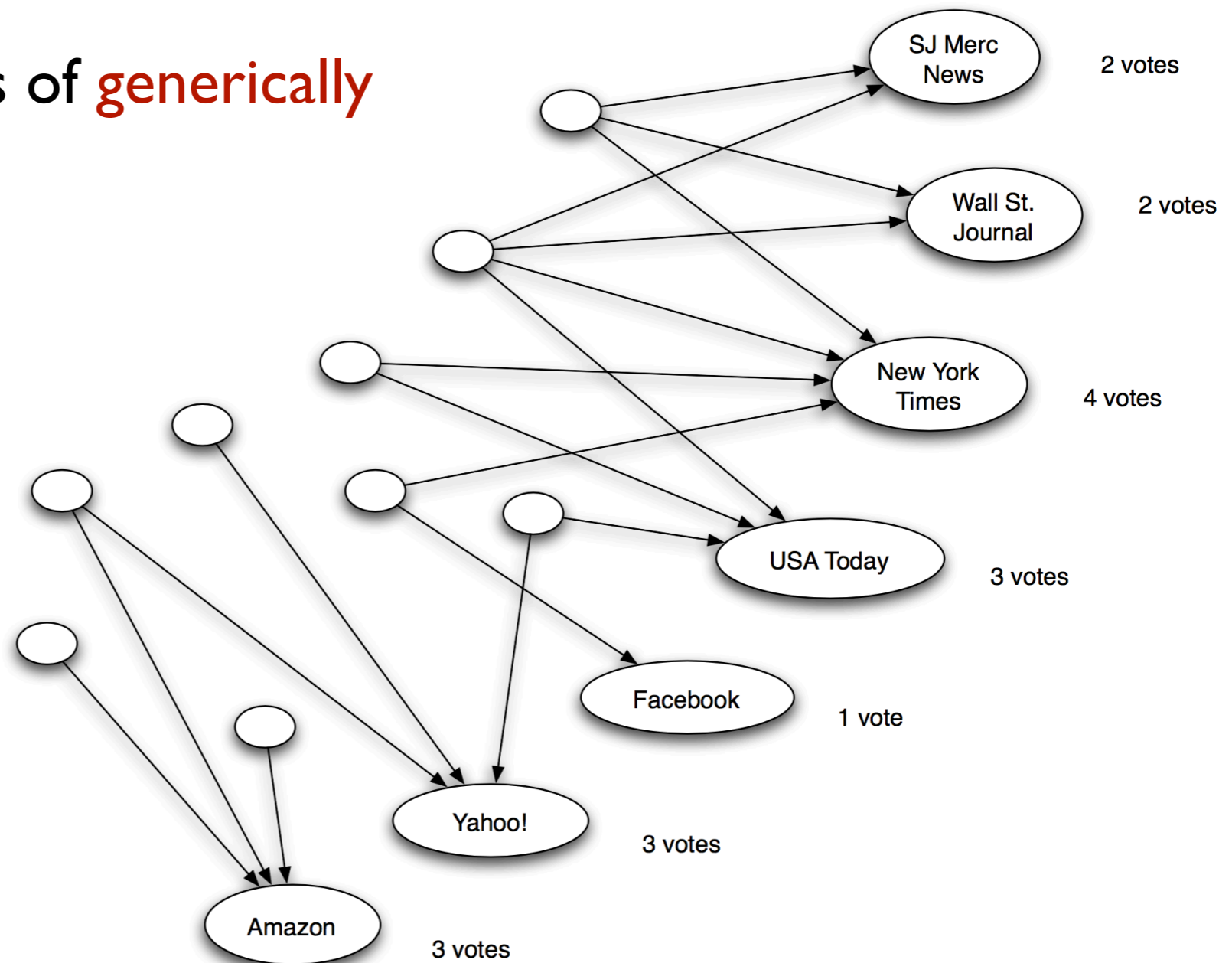
Personalised PageRank

Query: “newspaper”

No one right answer

Restrict to a relevant set and count the in-links

Lots of newspapers, but lots of **generically popular** sites too

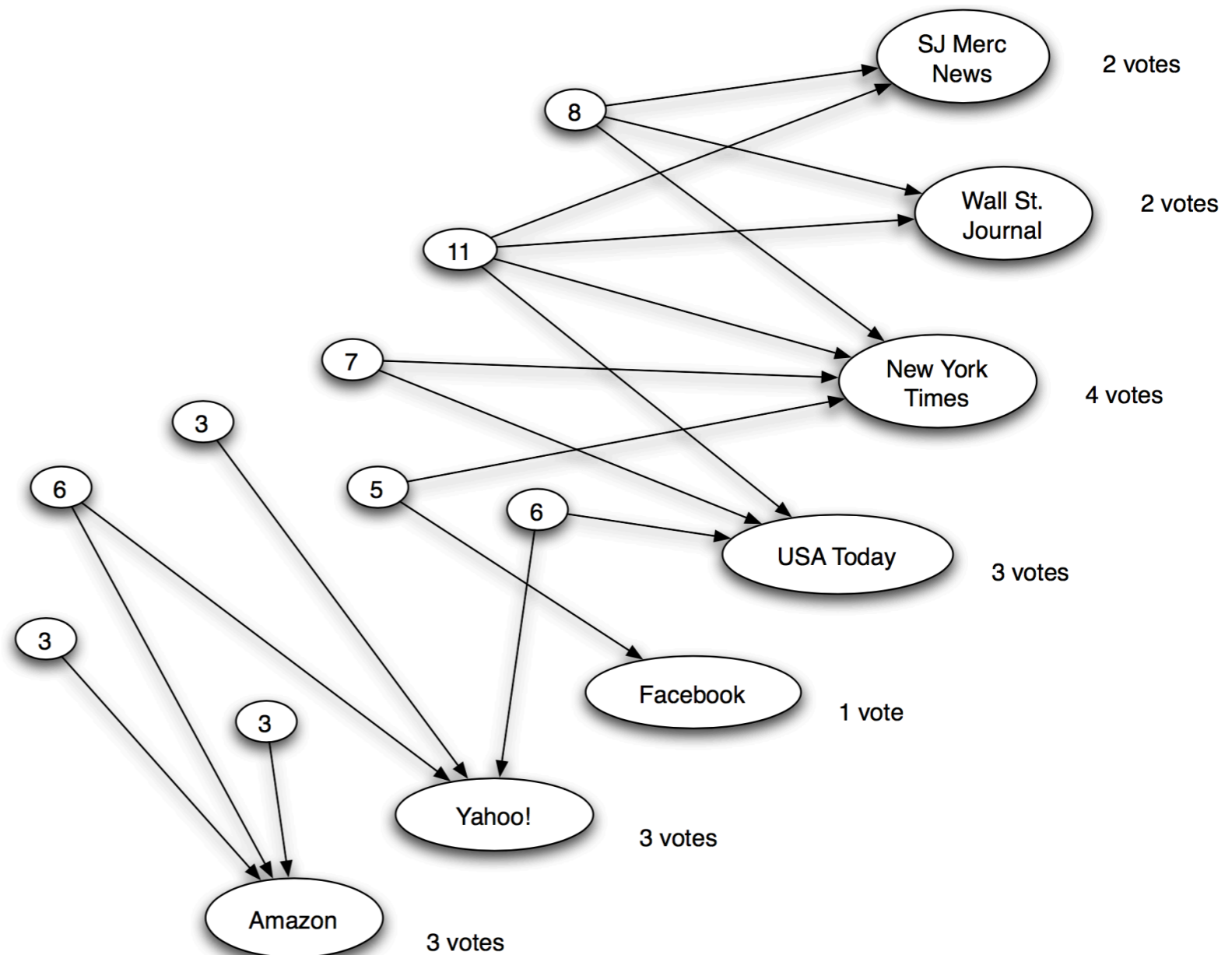


(Note this is idealized example. In reality the graph is not bipartite)

What about the lists?

Nodes on the left: some are better than others!

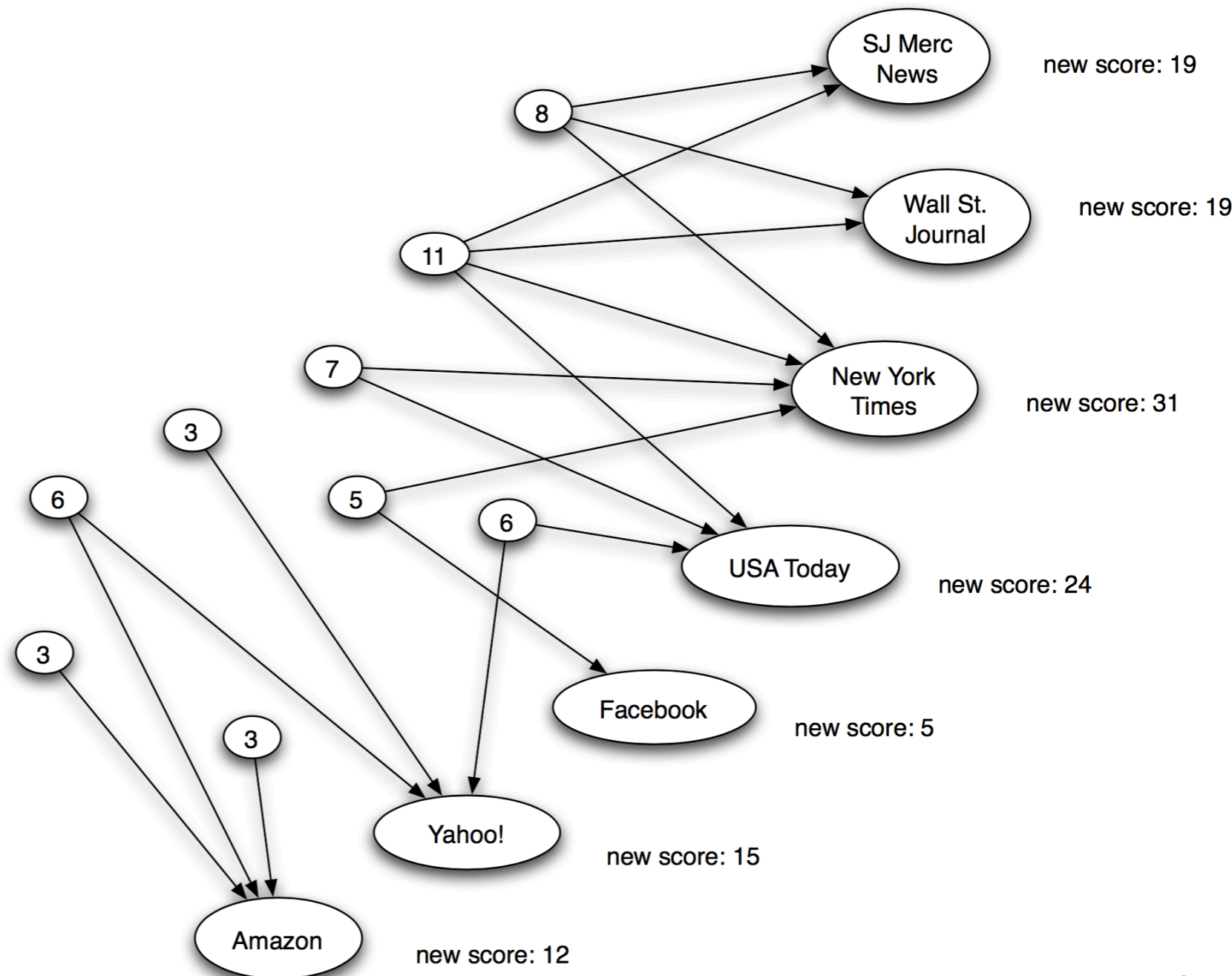
Value of a list = sum of votes received by the pages it links to



But wait! Reweight the pages

Linked to by good lists? Should be worth more

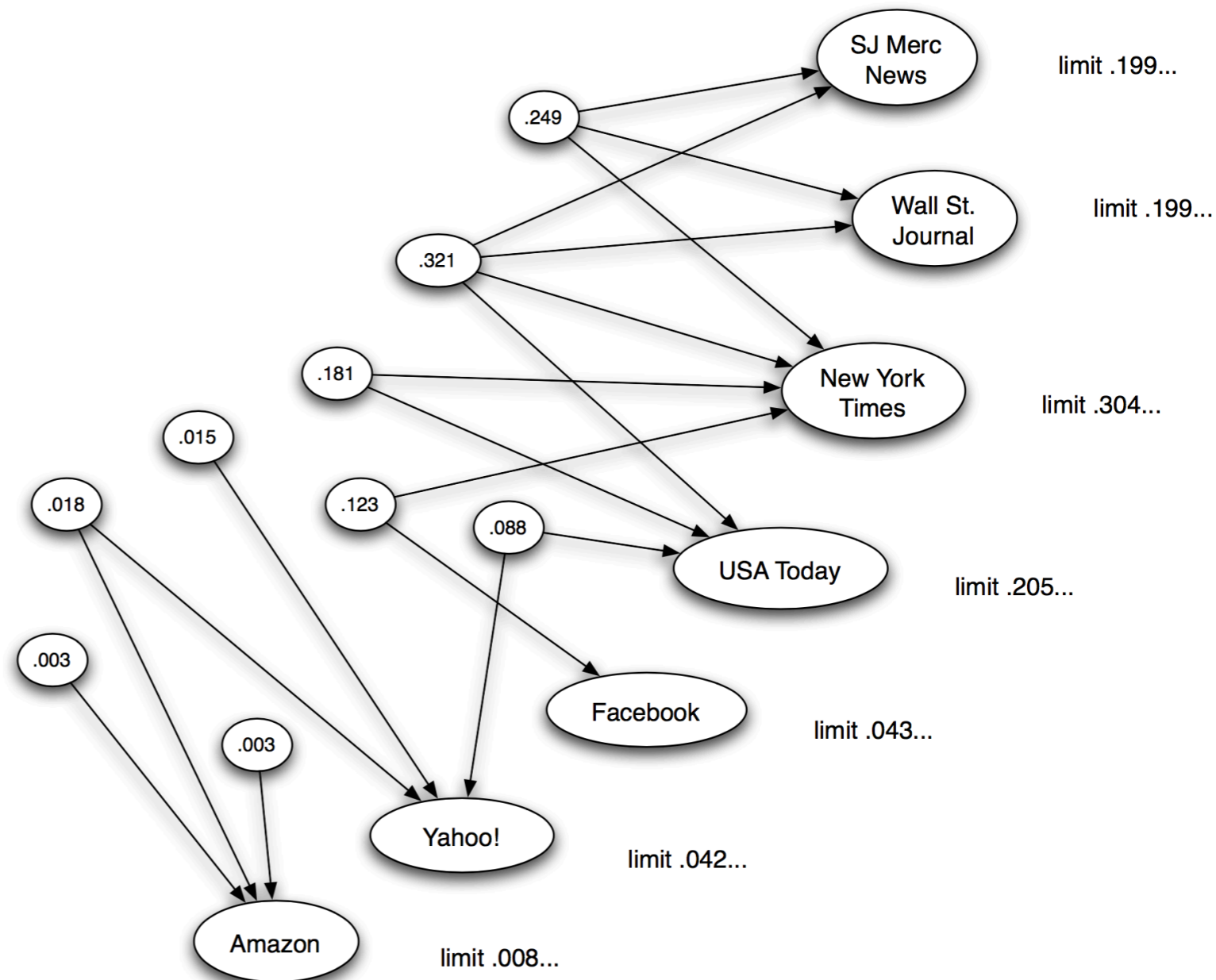
Value of a page: sum of list values of lists that linked to it



Principle of Iterated Improvement

Why stop there?

We can keep doing this over and over and get better list scores and page scores

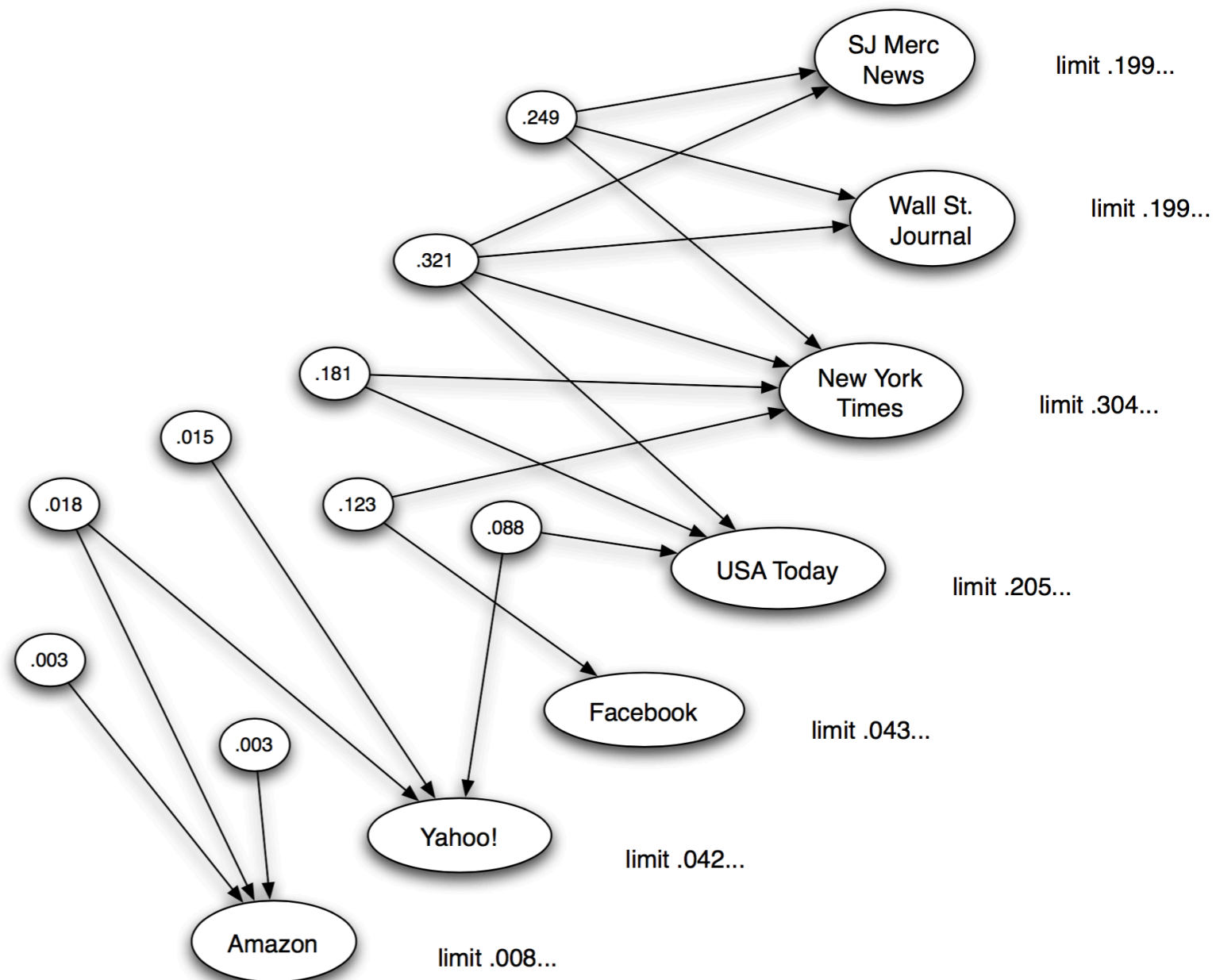


Principle of Iterated Improvement

Hubs and Authorities

Hubs: pages that are “lists” of links that link to good stuff

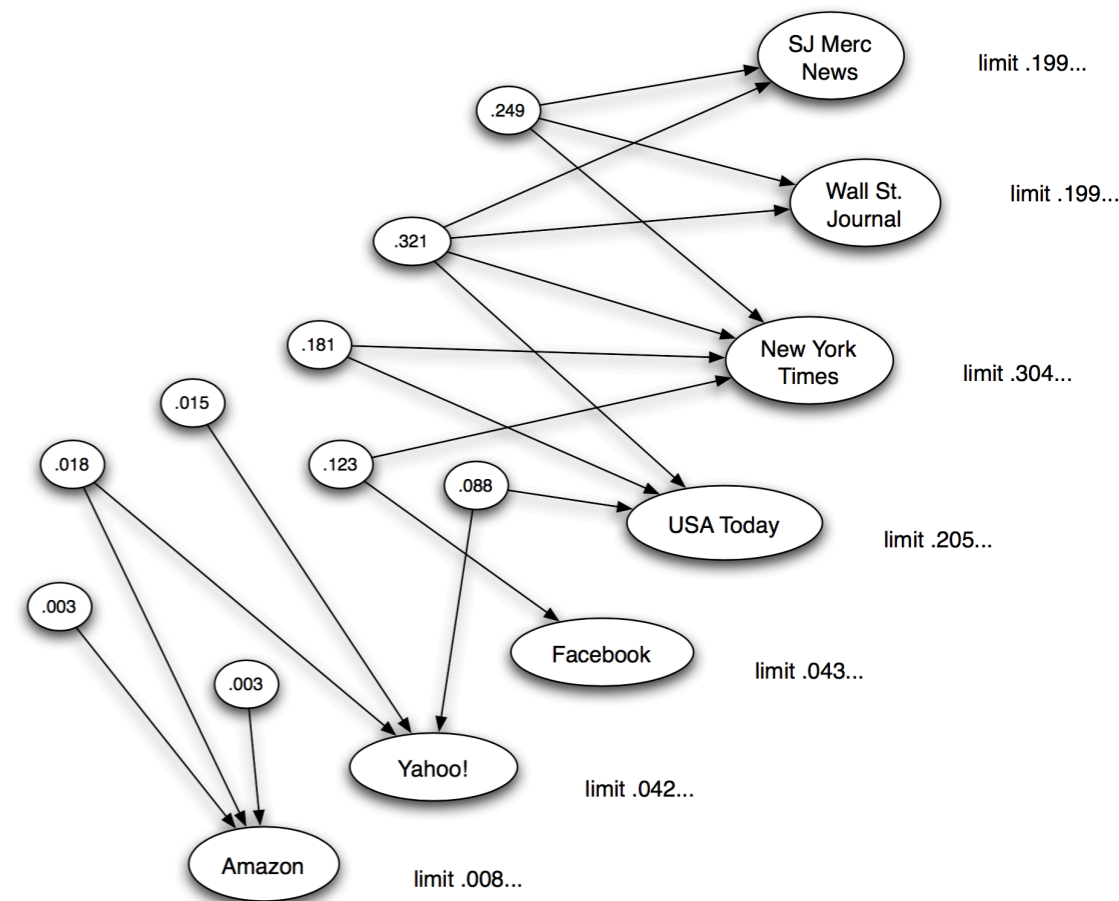
Authorities: pages that are good, authoritative... and linked to by good hubs



Principle of Iterated Improvement

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

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



Link Analysis: summary so far

Goal (back to the newspaper example):

Don't just find newspapers. Find “experts” – pages that link in a coordinated way to good newspapers

Idea: **Links as votes**

Hubs and Authorities

Each page has **2** scores:

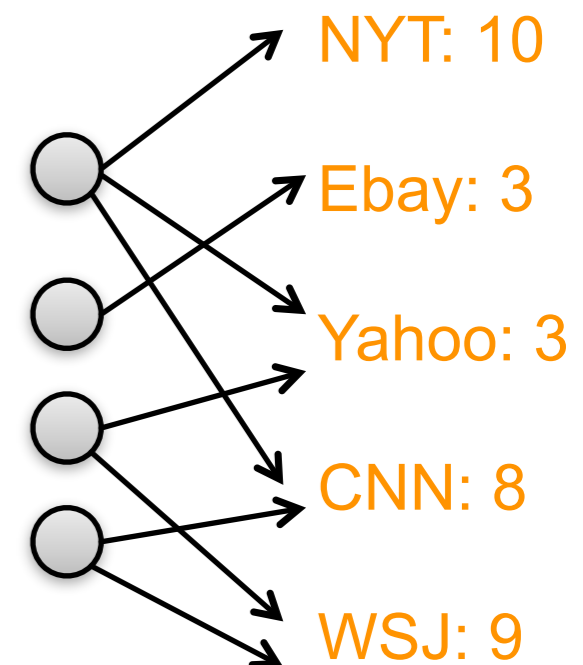
Quality as an expert (**hub**):

Total sum of votes of pages pointed to

Quality as an content (**authority**):

Total sum of votes of experts

Principle of repeated improvement



Hubs and Authorities

Interesting pages fall into **two classes**:

1. **Authorities** are pages containing useful information

Newspaper home pages

Course home pages

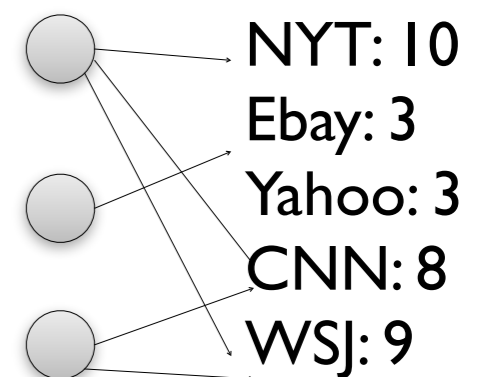
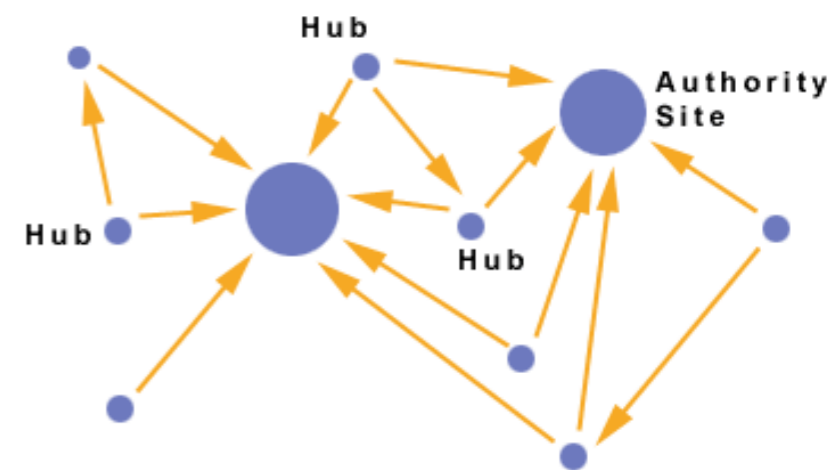
Home pages of auto manufacturers

2. **Hubs** are pages that link to authorities

List of newspapers

Course bulletin

List of U.S. auto manufacturers



Mutually Recursive Definition

- A good hub links to many good authorities
- A good authority is linked from many good hubs
 - Note a self-reinforcing recursive definition
- Model using two scores for each node:
 - Hub score and Authority score
 - Represented as vectors \mathbf{h} and \mathbf{a} , where the i -th element is the hub/authority score of the i -th node

Hubs and Authorities

Each page has a hub score h_i and an authority score a_i

HITS algorithm:

1. **Initialize** all scores to 1
2. Perform a sequence of **hub-authority updates**:
 - First apply **Authority Update Rule**
 - Then apply **Hub Update Rule**
3. **Normalize** (divide authority scores by sum over a_i 's and same for hubs)

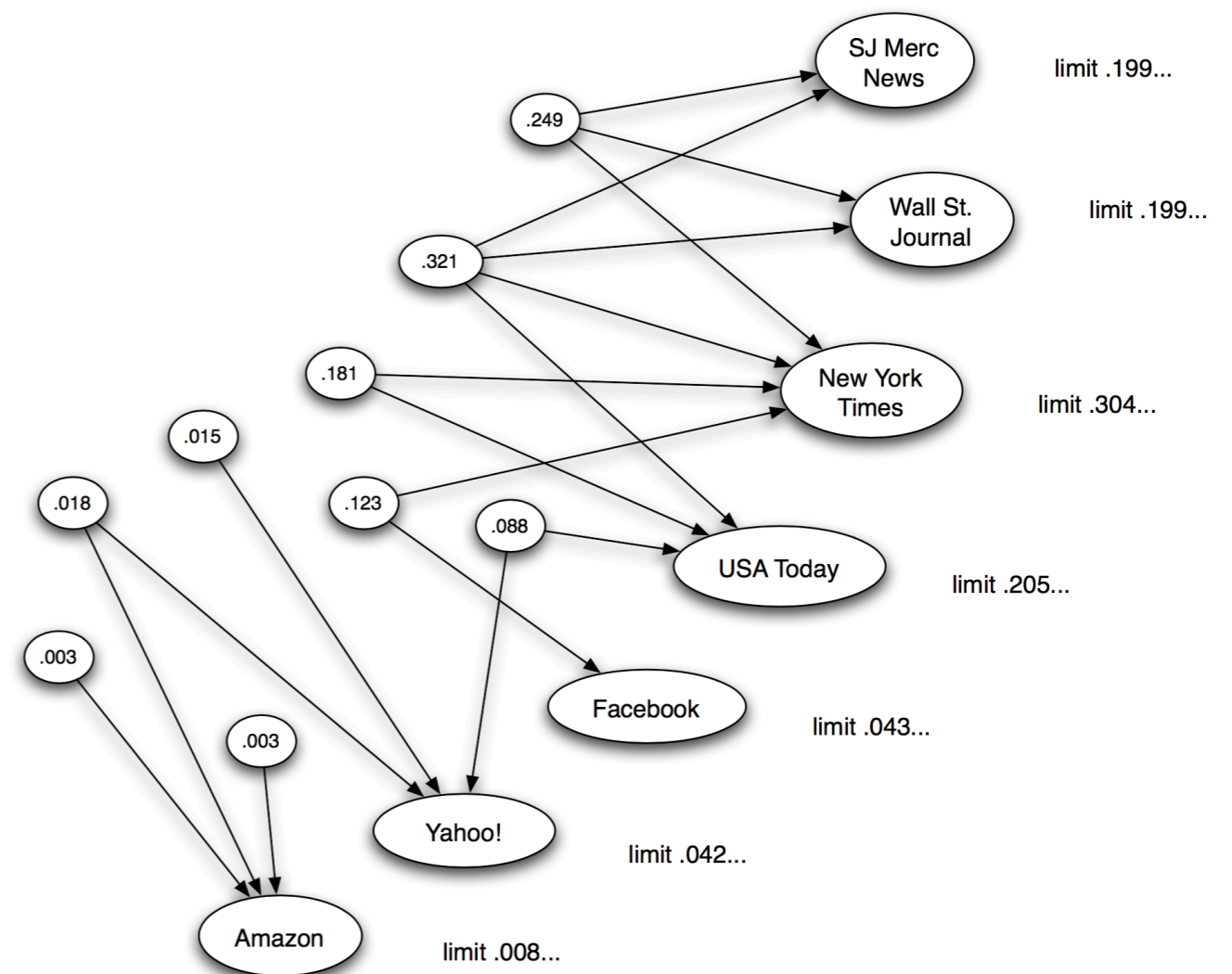
(We normalize since the numbers get very big,
and we only care about the relative sizes)

Hubs and Authorities: What happens?

What happens after a lot of steps?

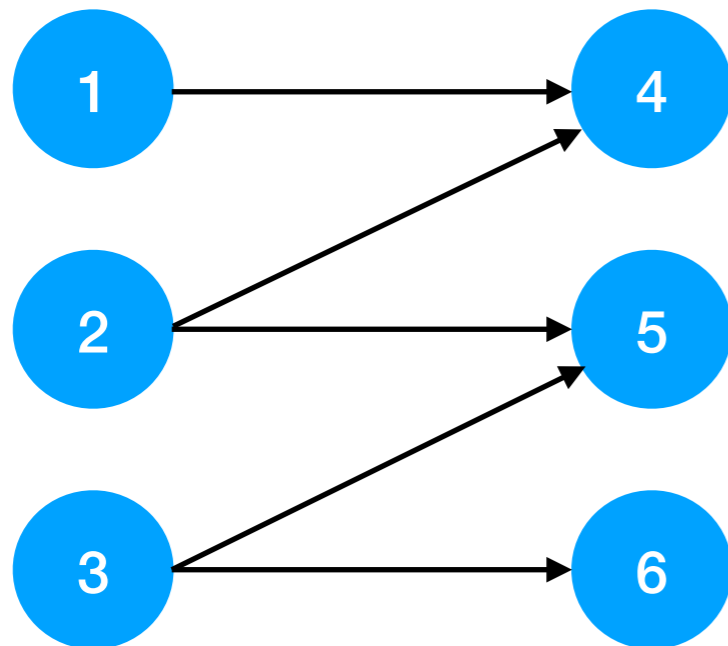
No matter what the starting scores are, **it always converges to the same hub and authority scores!**

Really a property of the link structure



Hubs and Authorities: Example

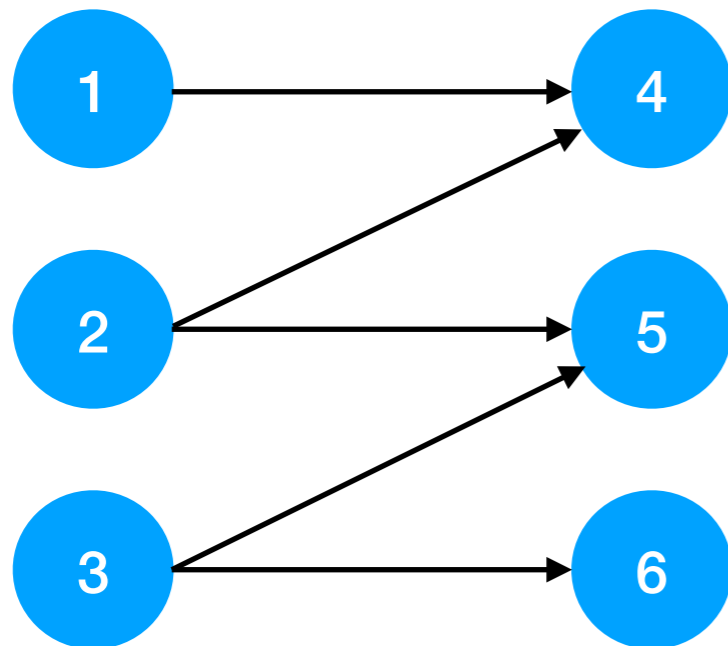
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1				
2	1				
3	1				
4	1				
5	1				
6	1				

Hubs and Authorities: Example

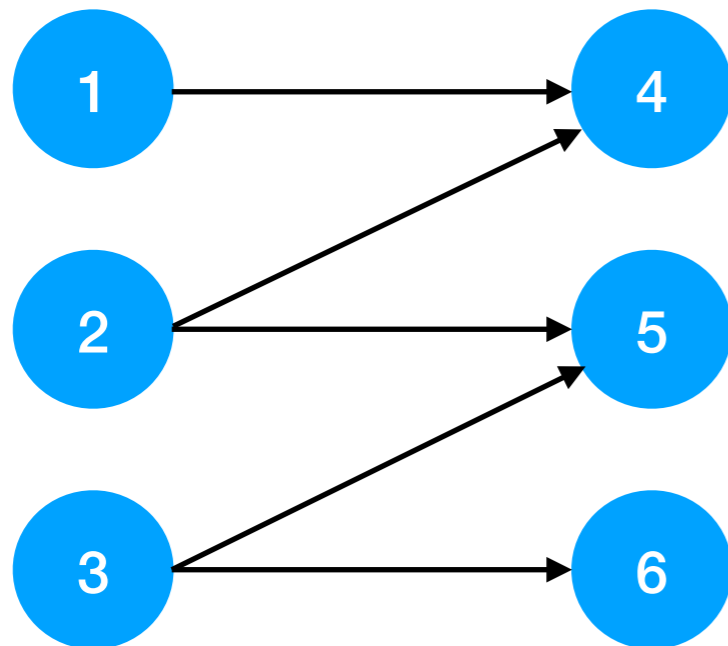
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0			
2	1	0			
3	1	0			
4	1	2			
5	1	2			
6	1	1			

Hubs and Authorities: Example

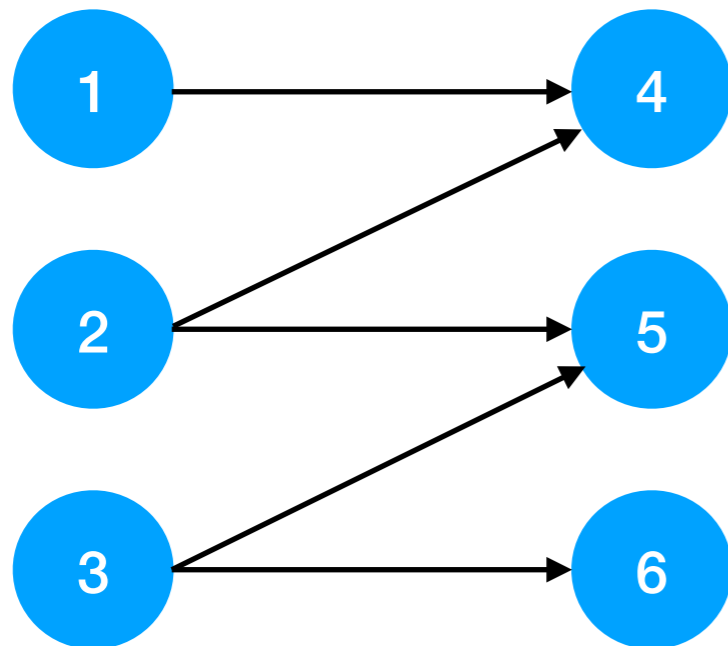
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0			
2	1	0			
3	1	0			
4	1	$2/5$			
5	1	$2/5$			
6	1	$1/5$			

Hubs and Authorities: Example

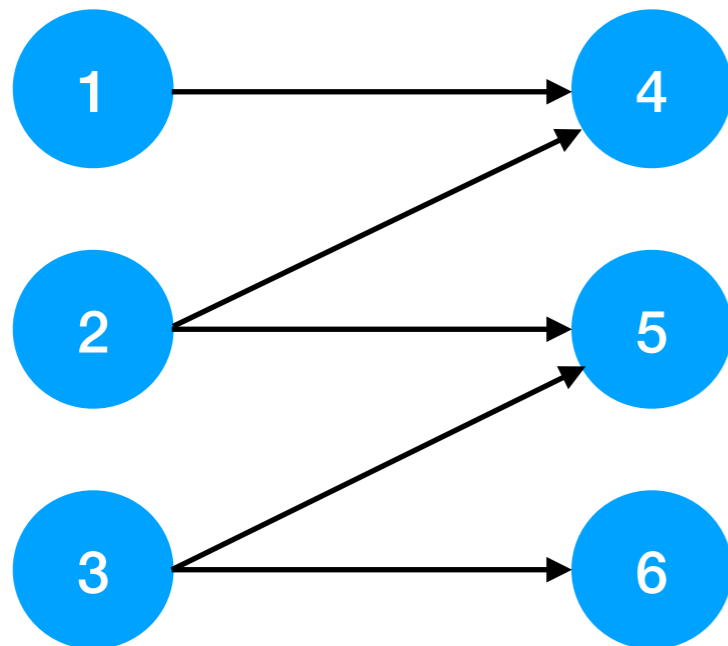
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	$2/5$		
2	1	0	$4/5$		
3	1	0	$3/5$		
4	1	$2/5$	0		
5	1	$2/5$	0		
6	1	$1/5$	0		

Hubs and Authorities: Example

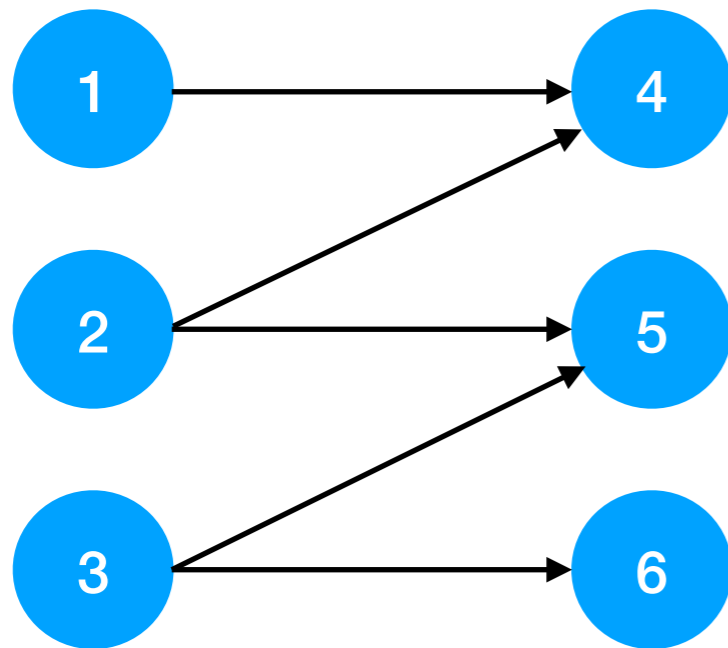
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	$(2/5)/(9/5)$		
2	1	0	$(4/5)/(9/5)$		
3	1	0	$(3/5)/(9/5)$		
4	1	$2/5$	0		
5	1	$2/5$	0		
6	1	$1/5$	0		

Hubs and Authorities: Example

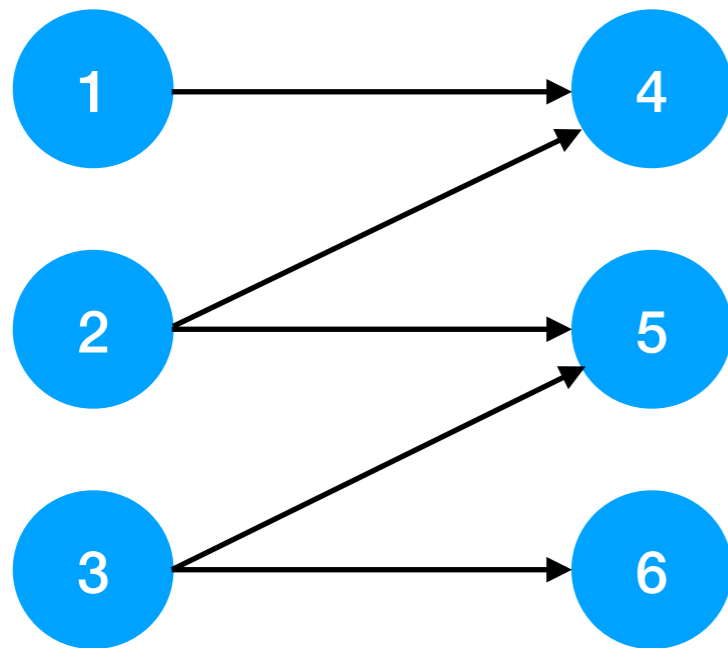
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	2/9		
2	1	0	4/9		
3	1	0	3/9		
4	1	2/5	0		
5	1	2/5	0		
6	1	1/5	0		

Hubs and Authorities: Example

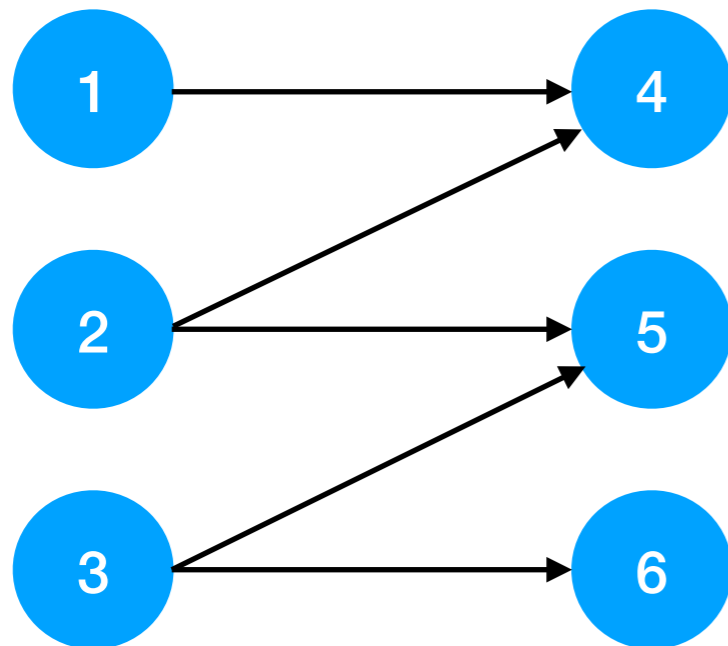
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	2/9	0	
2	1	0	4/9	0	
3	1	0	3/9	0	
4	1	2/5	0	6/9	
5	1	2/5	0	7/9	
6	1	1/5	0	3/9	

Hubs and Authorities: Example

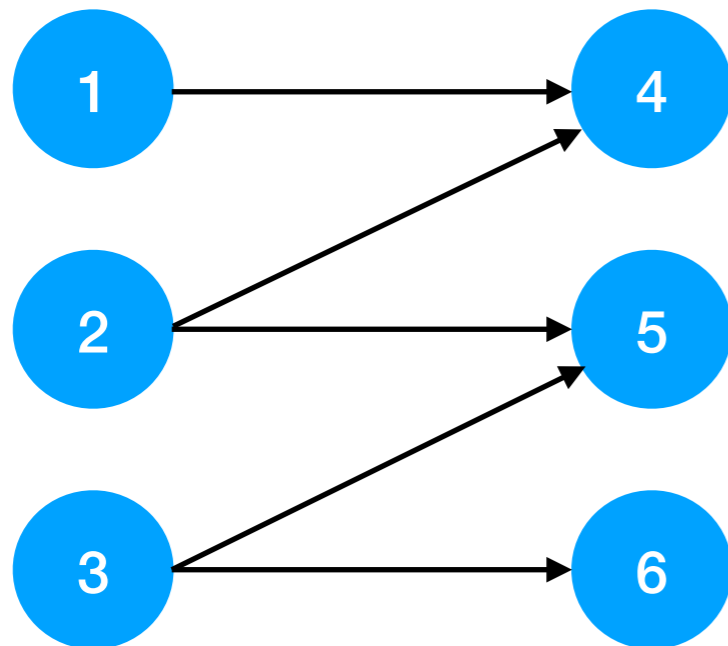
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	$2/9$	0	
2	1	0	$4/9$	0	
3	1	0	$3/9$	0	
4	1	$2/5$	0	$(6/9)/(16/9)$	
5	1	$2/5$	0	$(7/9)/(16/9)$	
6	1	$1/5$	0	$(3/9)/(16/9)$	

Hubs and Authorities: Example

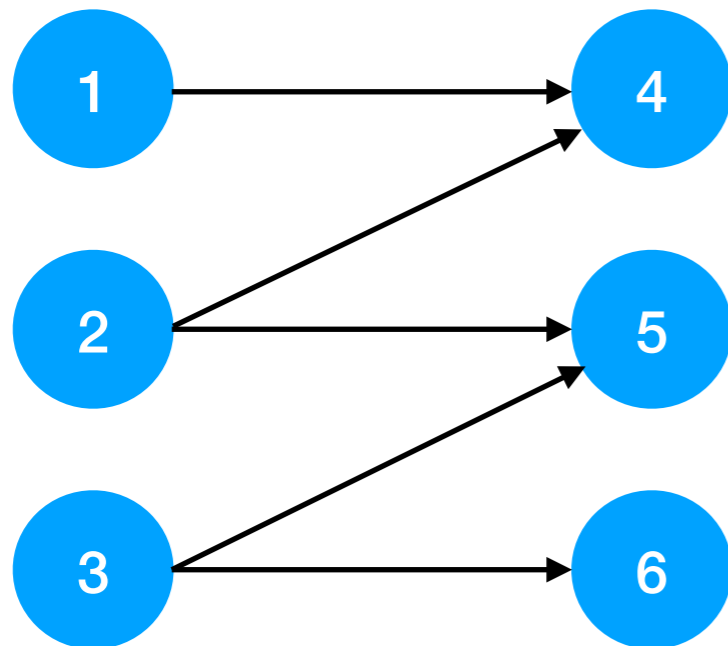
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	2/9	0	
2	1	0	4/9	0	
3	1	0	3/9	0	
4	1	2/5	0	6/16	
5	1	2/5	0	7/16	
6	1	1/5	0	3/16	

Hubs and Authorities: Example

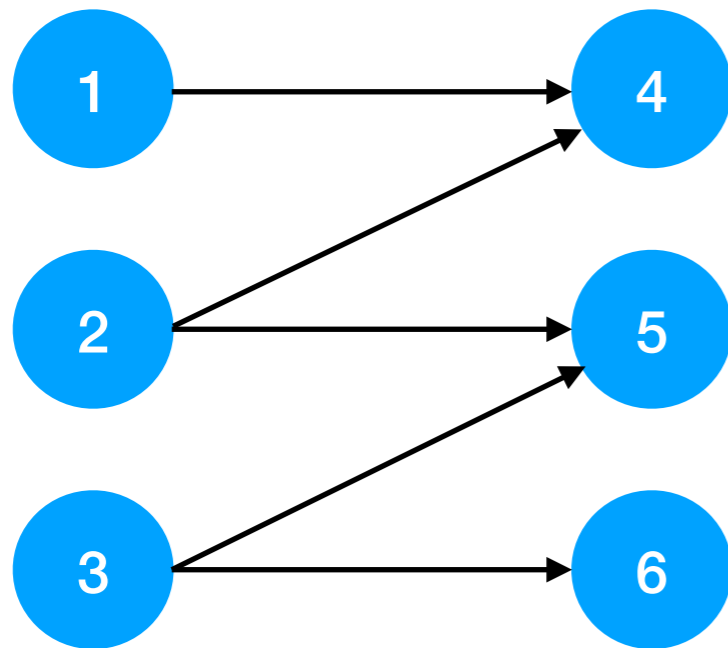
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	2/9	0	6/16
2	1	0	4/9	0	13/16
3	1	0	3/9	0	10/16
4	1	2/5	0	6/16	0
5	1	2/5	0	7/16	0
6	1	1/5	0	3/16	0

Hubs and Authorities: Example

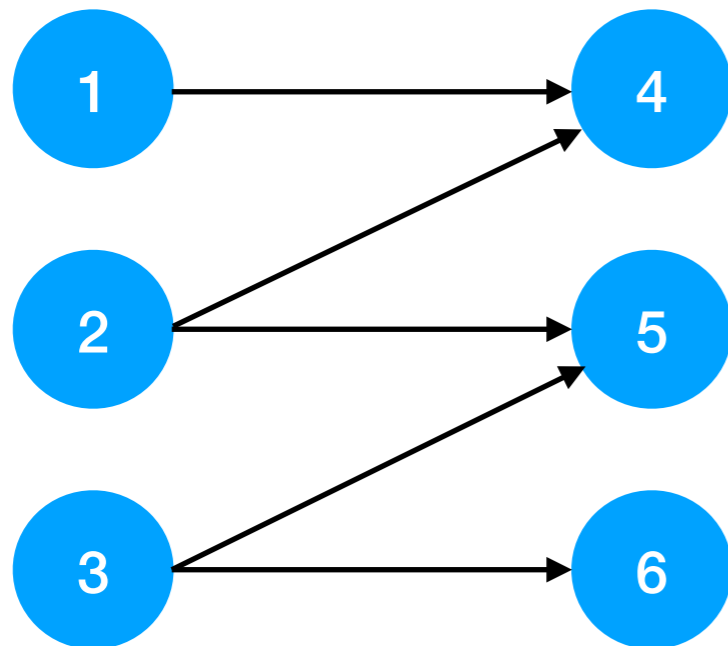
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	2/9	0	$(6/16)/(29/16)$
2	1	0	4/9	0	$(13/16)/(29/16)$
3	1	0	3/9	0	$(10/16)/(29/16)$
4	1	2/5	0	6/16	0
5	1	2/5	0	7/16	0
6	1	1/5	0	3/16	0

Hubs and Authorities: Example

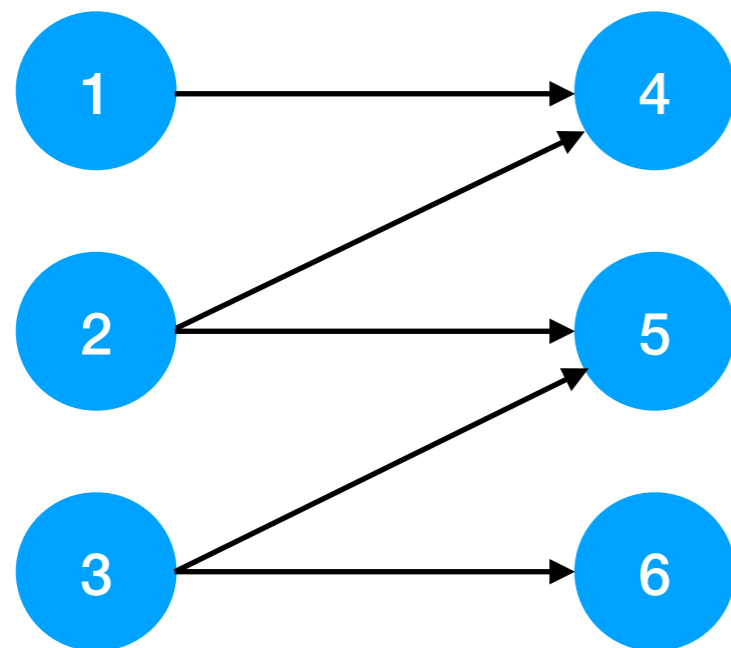
Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$
1	1	0	2/9	0	6/29
2	1	0	4/9	0	13/29
3	1	0	3/9	0	10/29
4	1	2/5	0	6/16	0
5	1	2/5	0	7/16	0
6	1	1/5	0	3/16	0

Hubs and Authorities: Example

Apply 2 rounds of hub and authority update steps on the graph below:



Node	$h^{<0>}$	$a^{<1>}$	$h^{<1>}$	$a^{<2>}$	$h^{<2>}$...	$a^{<*>}$	$h^{<*>}$
1	1	0	2/9	0	6/29	...	0	0.198
2	1	0	4/9	0	13/29	...	0	0.445
3	1	0	3/9	0	10/29	...	0	0.357
4	1	2/5	0	6/16	0	...	0.357	0
5	1	2/5	0	7/16	0	...	0.445	0
6	1	1/5	0	3/16	0	...	0.198	0

Note: in this example, values are very close to convergence after only 2 steps

PageRank

Links as Votes

Hubs and Authorities works well for situations where pages can be strong signals of quality (endorsers) without themselves being endorsed

Think of things like commercial pages, where competitors are **unlikely to link to each other, but may be endorsed by similar hubs**

But in many situations, importance passes directly from one prominent page to another

Links as Votes

Still the same idea: Links as votes

Page is more important if it has more links

In-coming links? Out-going links?

Think of in-links as votes:

www.utoronto.ca has 23,400 in-links

www.random-shady-website-uhoh.com has 1 in-link

Are all in-links equal?

Links from important pages count more

You're important if important nodes vote for you

Recursive question!

How would you set up an update rule with only one measure of importance, not two?

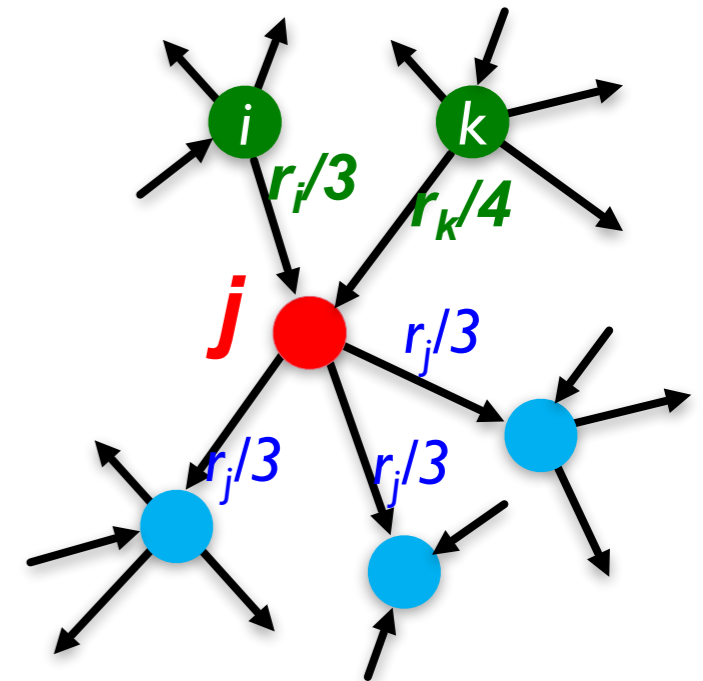
PageRank: The “Flow” Model

A “vote” from an important page is worth more:

Each link’s vote is proportional to the **importance** of its source page

If page i with importance r_i has d_i out-links, each link gets r_i / d_i votes

Page j ’s own importance r_j is the sum of the votes on its in-links



$$r_j = r_i/3 + r_k/4$$

Mental Model: PageRank as a Fluid

Think of PageRank as a “fluid” that circulates around the network, passing from node to node and pooling at the most important ones

PageRank Algorithm:

1. Initialize all nodes with $1/n$ PageRank
2. Perform k PageRank updates:

Basic PageRank Update Rule: Each page divides its current PageRank equally across its outgoing links. New PageRank is the sum of PR you receive.

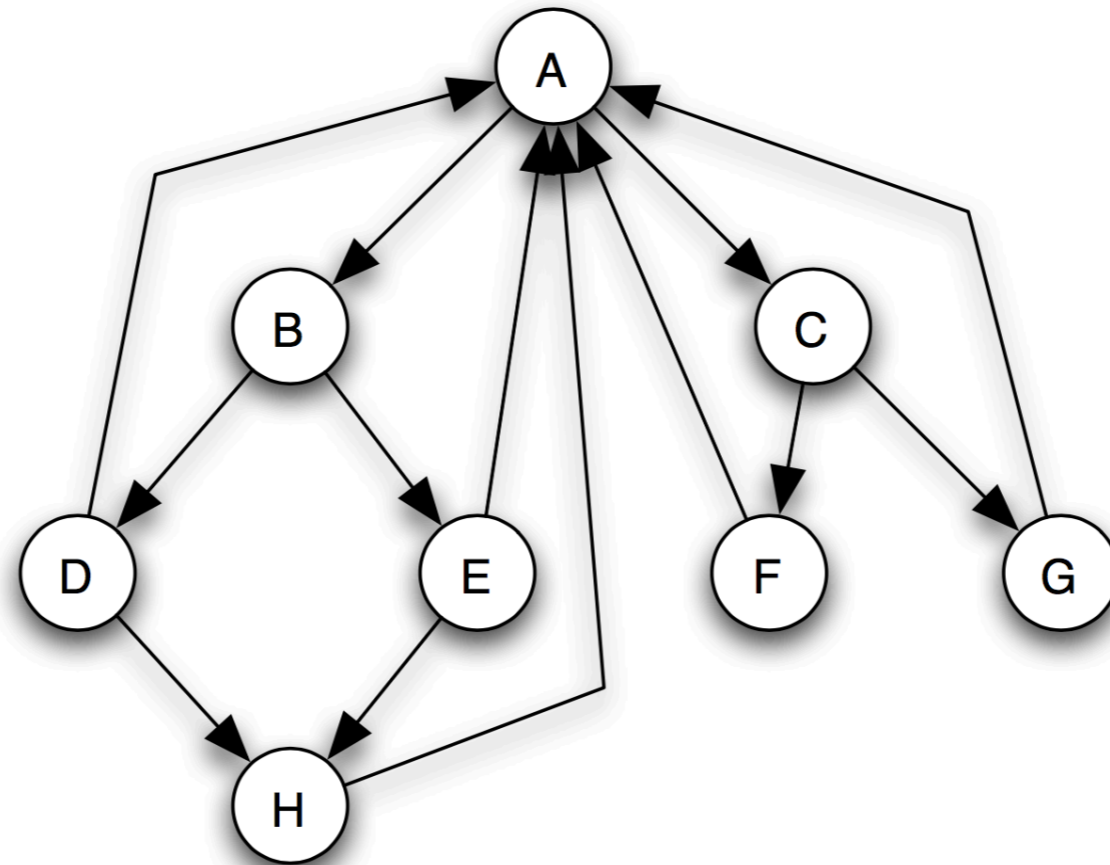
Page j 's PageRank Update equation:
$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$

Where d_i = out-degree of node i

PageRank: The “Flow” Model

Example: 8 nodes

Each starts with 1/8 PageRank

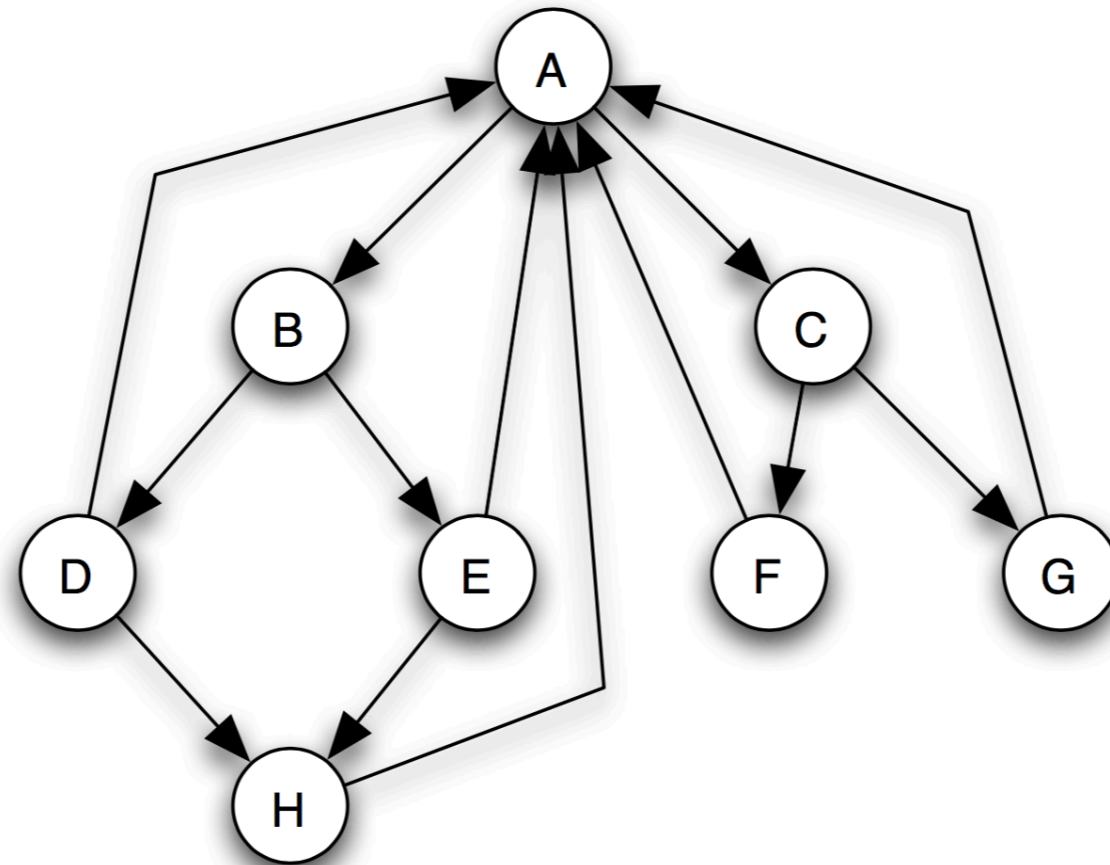


Step	A	B	C	D	E	F	G	H
1								
2								

PageRank: The “Flow” Model

Example: 8 nodes

Each starts with $1/8$ PageRank

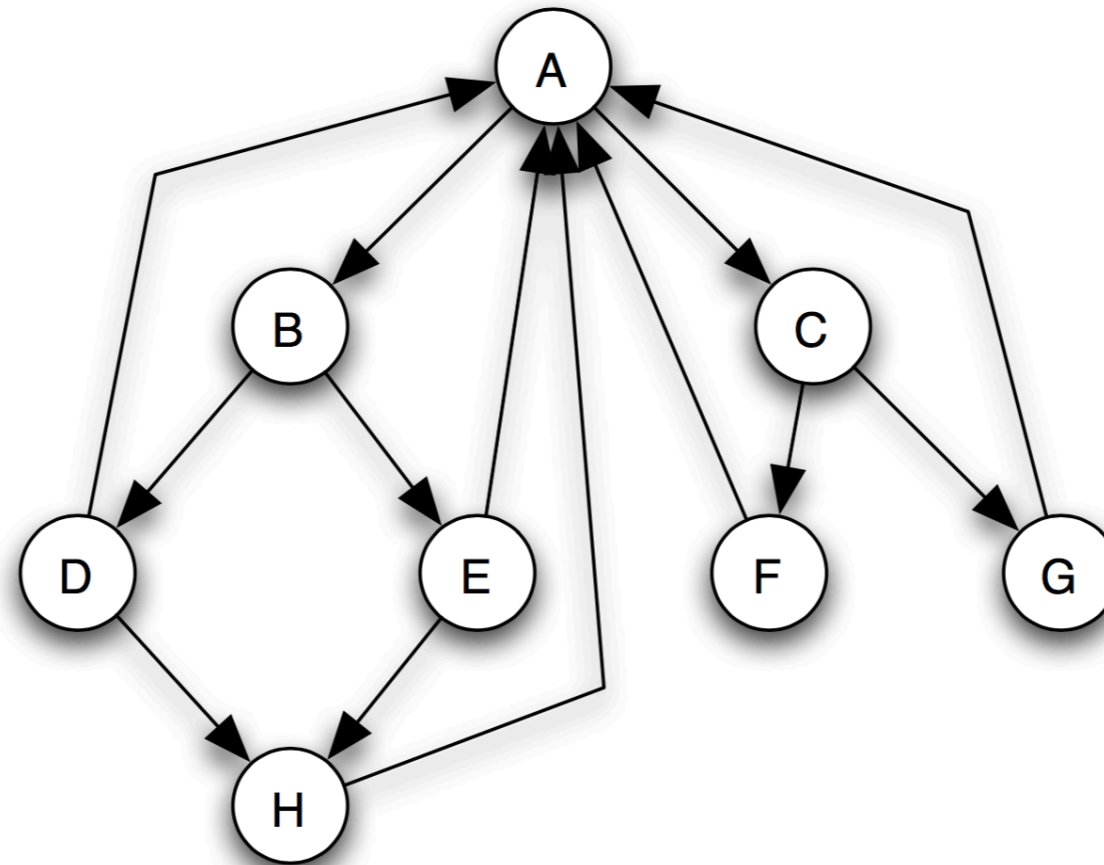


Step	A	B	C	D	E	F	G	H
1	$1/2$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/8$
2								

PageRank: The “Flow” Model

Example: 8 nodes

Each starts with $1/8$ PageRank



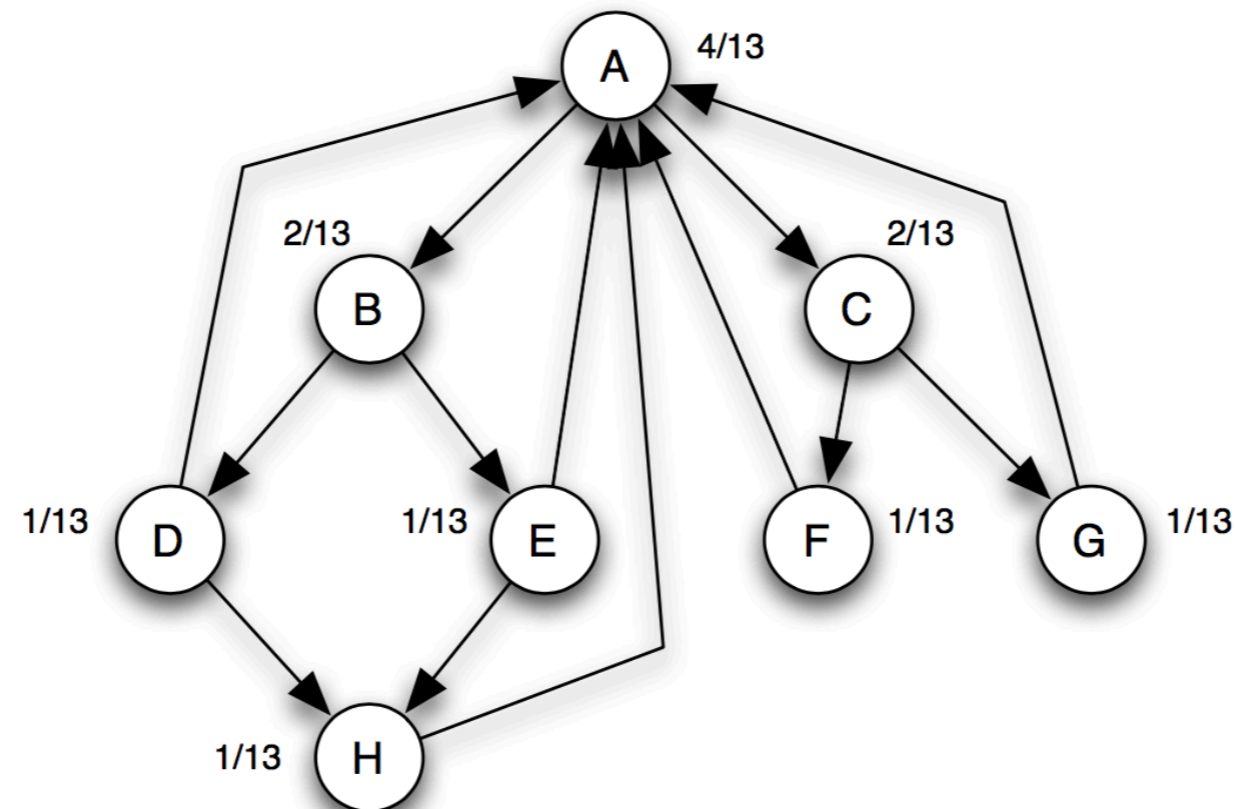
Step	A	B	C	D	E	F	G	H
1	$1/2$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/8$
2	$5/16$	$1/4$	$1/4$	$1/32$	$1/32$	$1/32$	$1/32$	$1/16$

PageRank: The “Flow” Model

Principle of Repeated Improvement!

As in H&A, this process **converges to limiting values**

In equilibrium, doing another PageRank Update **doesn't change anything**

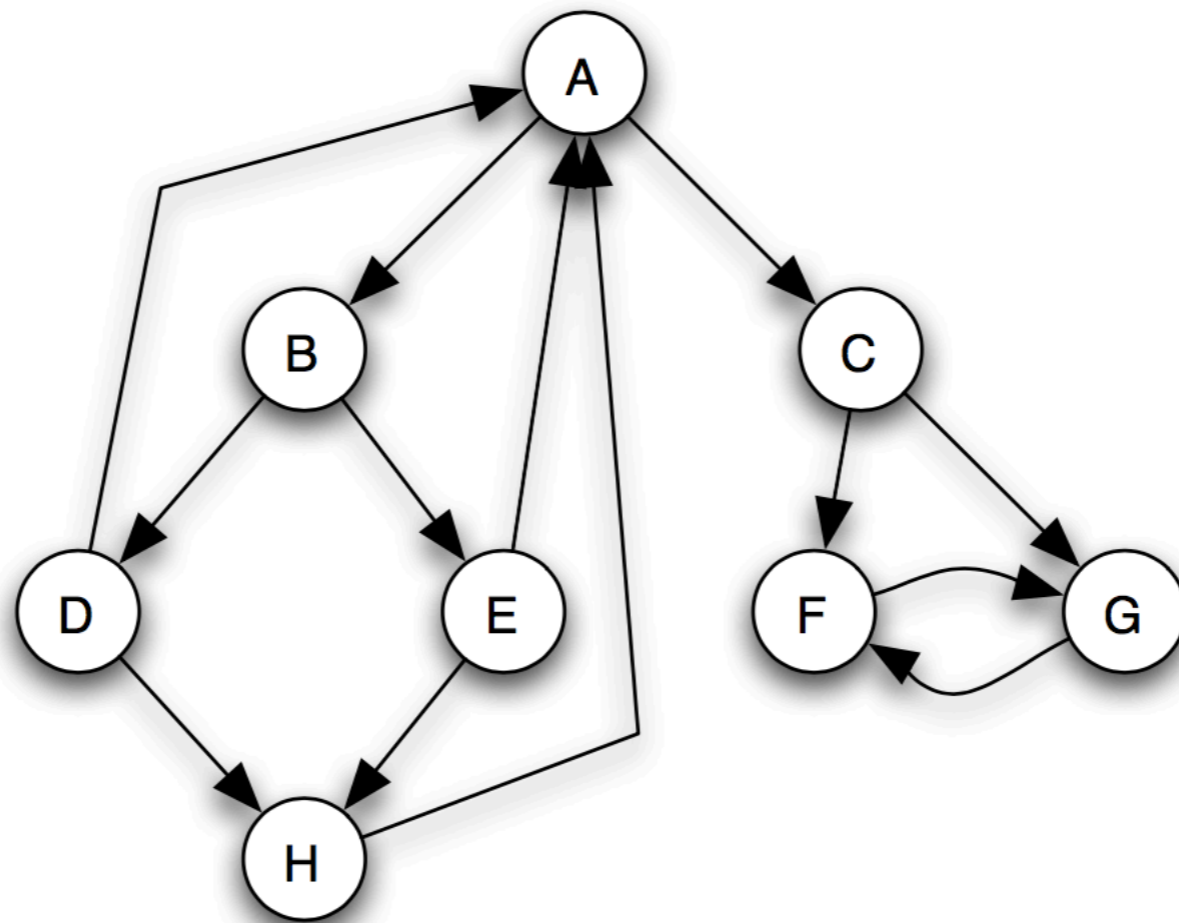


PageRank: A Problem

In real graph structures, PageRank can pool in the wrong places

Consider a slightly different graph:

What happens?

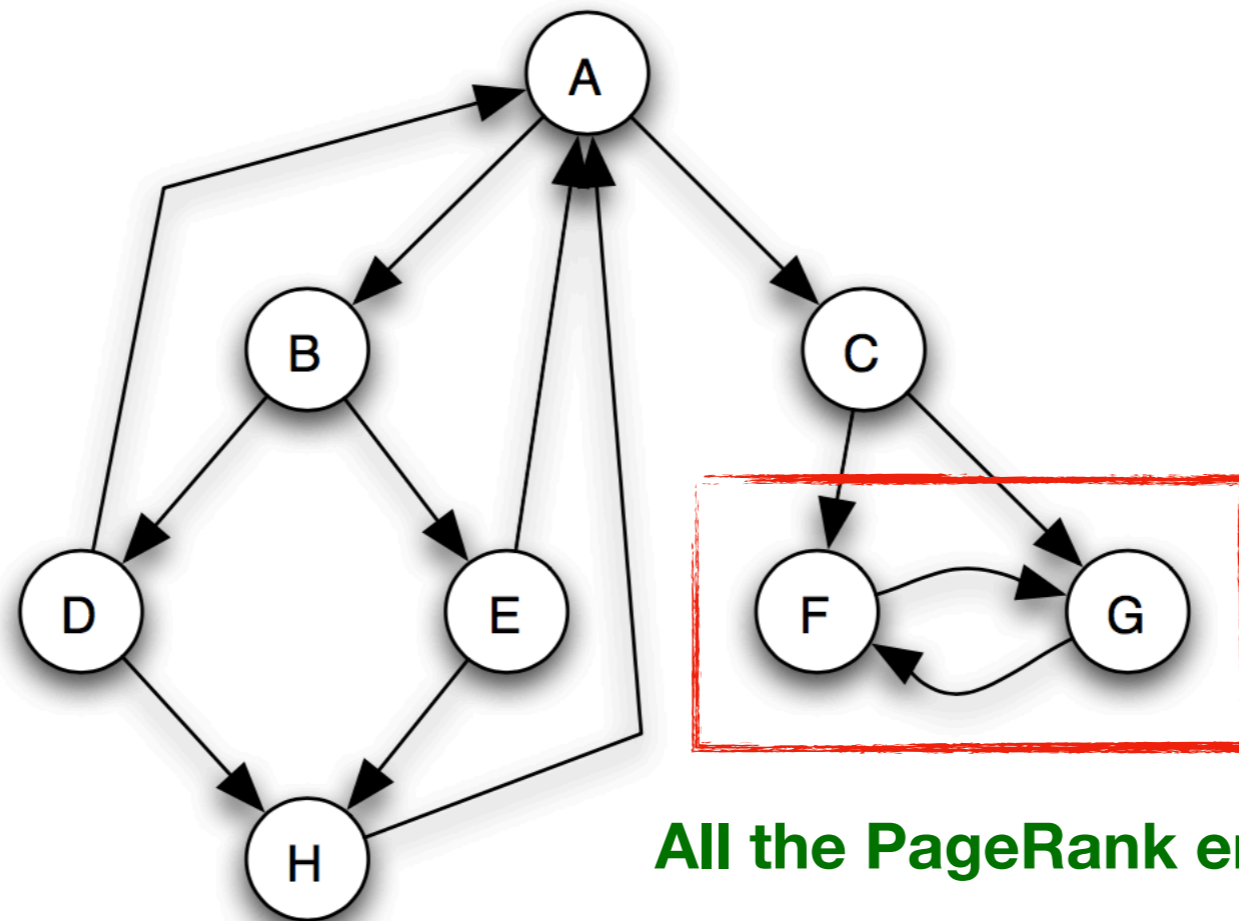


PageRank: A Problem

In real graph structures, PageRank can pool in the wrong places

Consider a slightly different graph:

What happens?



All the PageRank ends up here!

PageRank: A Solution

Scaled PageRank: only divide a fraction s of the PageRank among outgoing links

The rest gets **spread evenly over all nodes**

In effect we create a **complete graph**

Scaled PageRank Update Rule: First **apply** Basic PageRank Update Rule, **scale down** the values by s , then **divide the residual** $1-s$ units of PageRank equally: $(1-s)/n$ to each.

PageRank: A Solution

Scaled PageRank **converges** to a **unique set of equilibrium values** (but it depends on s)

This is the PageRank used in practice, with s chosen between 0.8 and 0.9.

PageRank: Random Surfer

Say a web surfer **navigates links randomly** (choosing each out-link with equal probability)

This is equivalent to the **Basic PageRank Update!**

Say he also jumps to a random node in the graph with probability $1-s$ (“Random Restarts”)

This is equivalent to the **Scaled PageRank Update**

One way to think about PageRank of a node: limiting probability that a random surfer ends up at that node

PageRank: Random Surfer

Claim: The probability of being at page X after k steps of this random walk is equal to the PageRank of X after k applications of the Basic PageRank Update rule.

The Random Walk: Walker chooses a starting node at random, then at each step picks one of the out-links of its current node uniformly at random.

Proof: Let $r_1, r_2, r_3, \dots, r_n$ represent probability of being at nodes $1, 2, 3, \dots, n$ in a given step of the random walk. Given these, what is probability of being at node j in next step?

- For each node i that links to j , $1/d_i$ prob of going to j
- Need to be at i for that to happen, so prob contribution is $r_i * 1/d_i$
- Summing over nodes, $r'_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$ **Same as before!**

Web Search in Practice

Link Analysis was the **basic revolutionary idea**

Still used today

But now, lots of other signals used for ranking

- Anchor text:

- “I am a student at University of Toronto”

- Can include in link analysis framework (give more weight to highly relevant anchor text)

- Click data

- If everyone clicks on the second link, rank it first

PageRank Applications

PageRank Application: Online Dating

How do people pursue romantic partners online?

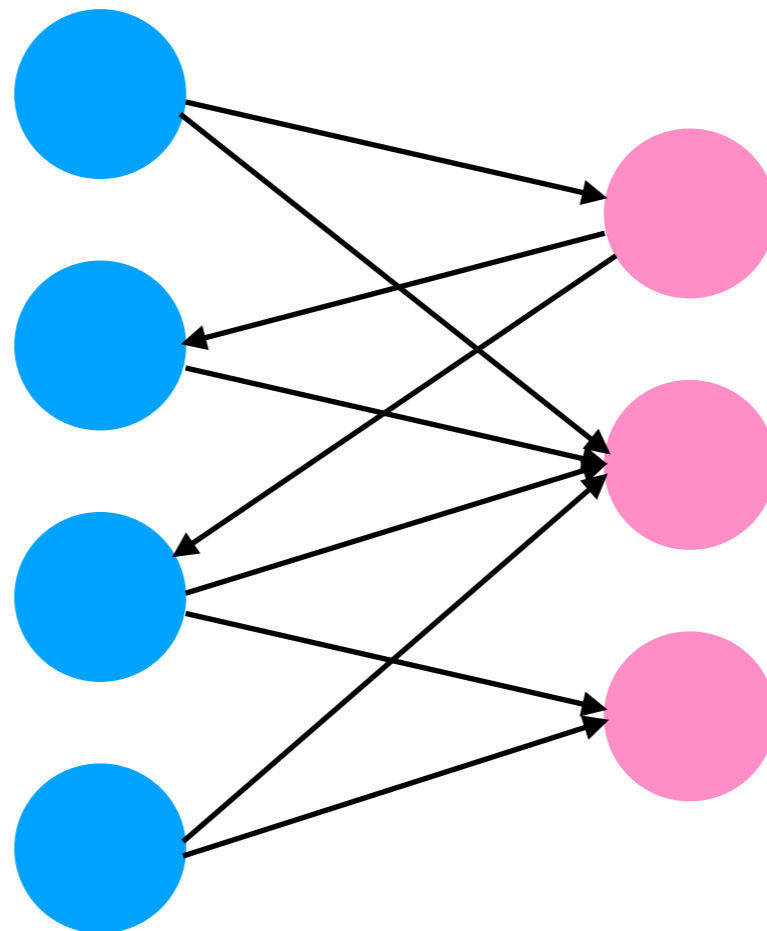
In particular, how does **desirability** play into how people look for partners?

How do you define “**desirability**”?

All sorts of problems with trying to infer it from someone’s profile page

PageRank Application: Online Dating

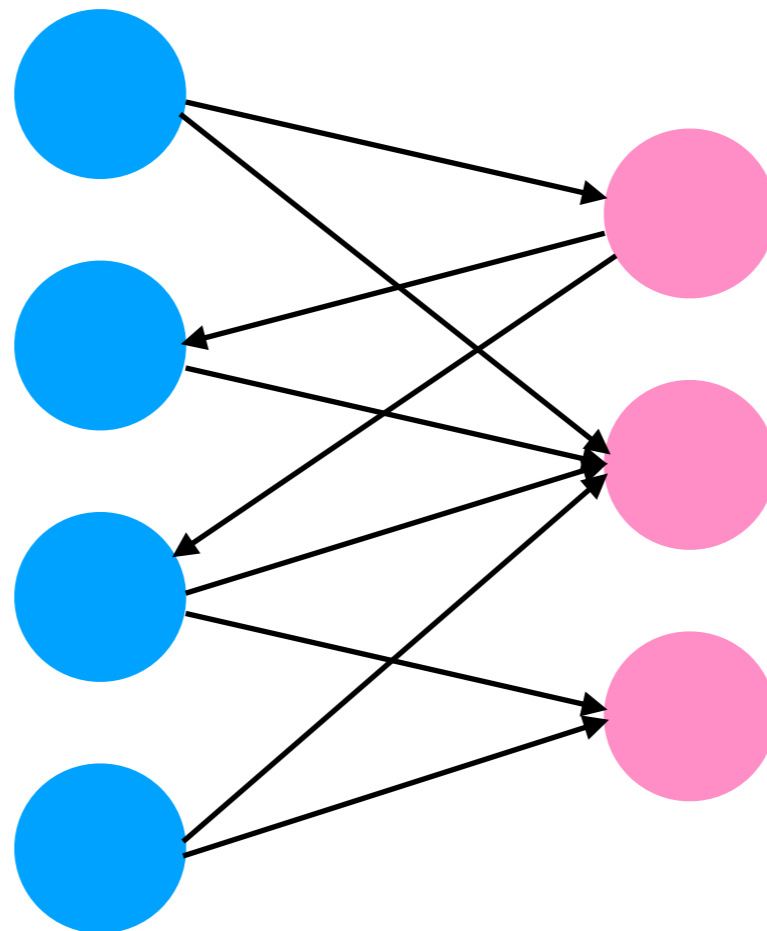
Consider the network of **messages exchanged** [Bruch and Newman, 2018]



N.B. The data in this paper is on heterosexual dating only.

PageRank Application: Online Dating

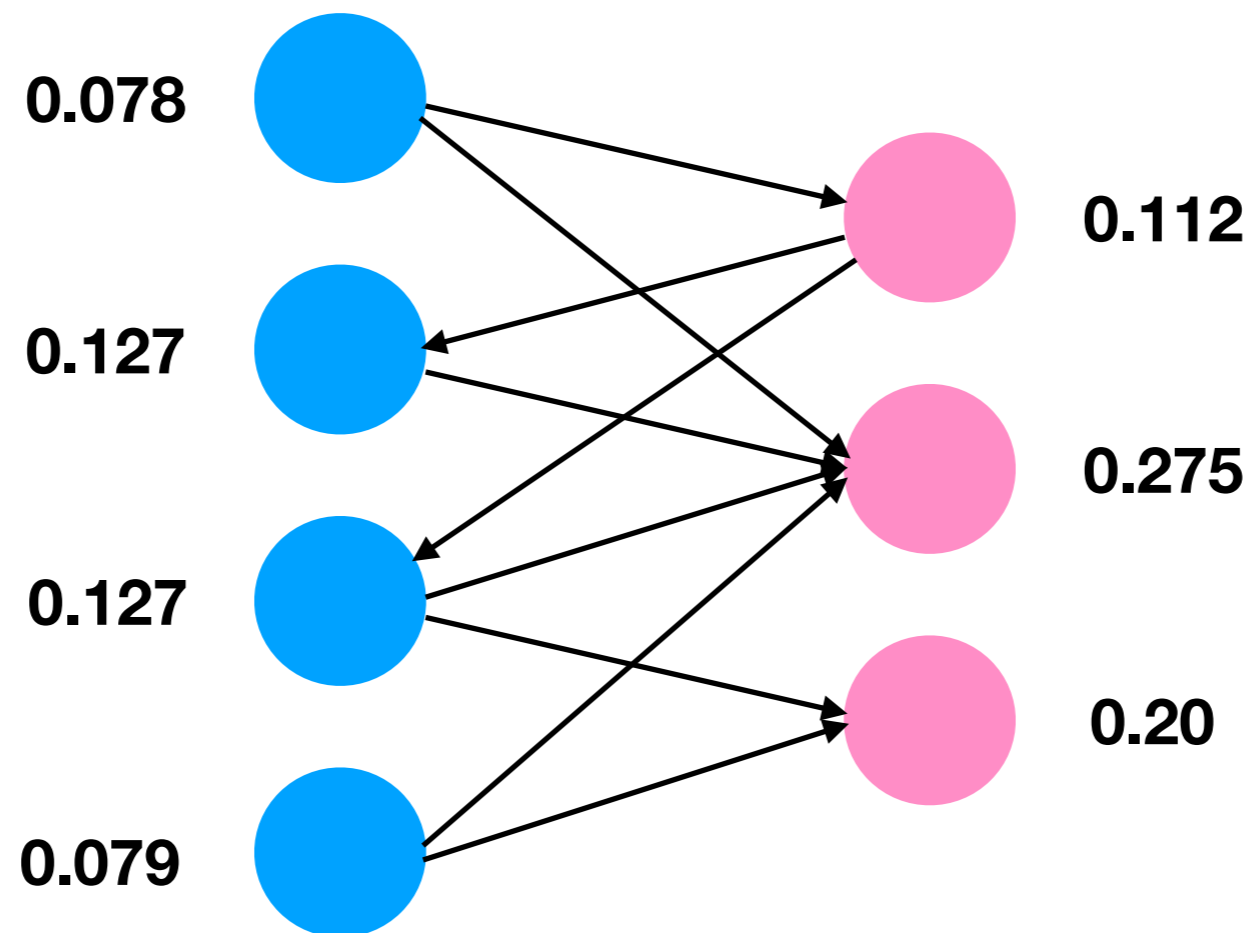
As before, there is **significant information** in the links!



Desirability: one is desirable if they are contacted a lot, and if they are contacted by other desirable people

PageRank Application: Online Dating

Idea: Apply PageRank to the initial-contact graph

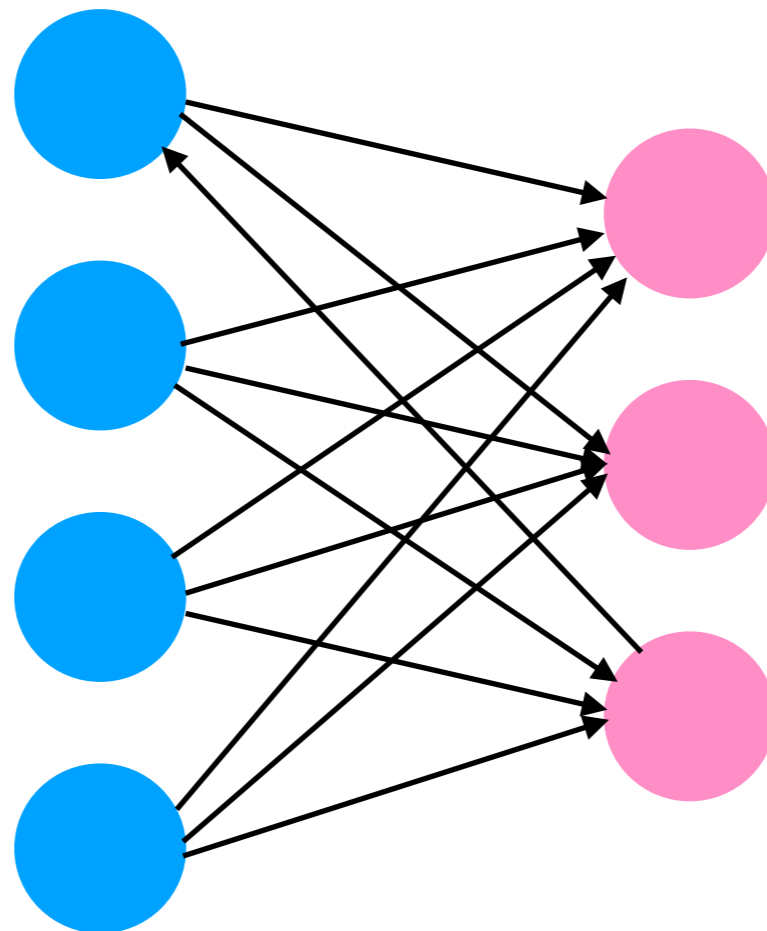


This gives a ranking over people using all of the information in their messaging behaviour

PageRank Application: Online Dating

Problem: Real-life initial-contact graphs look more like this:

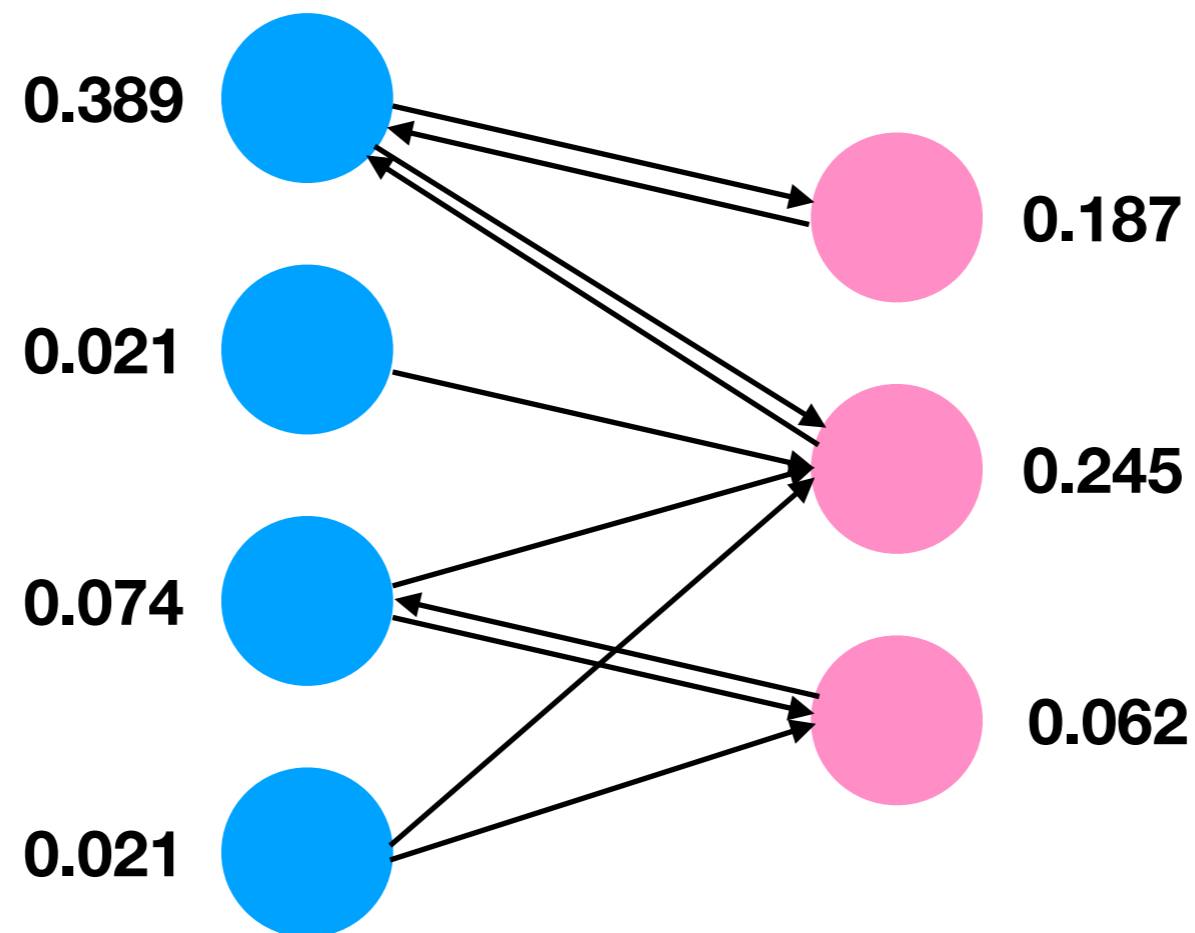
Males do most of the
initial messaging
(>80% in their data)



If you choose this representation (just the initial-contact graph), you don't get enough information about the females' preferences

PageRank Application: Online Dating

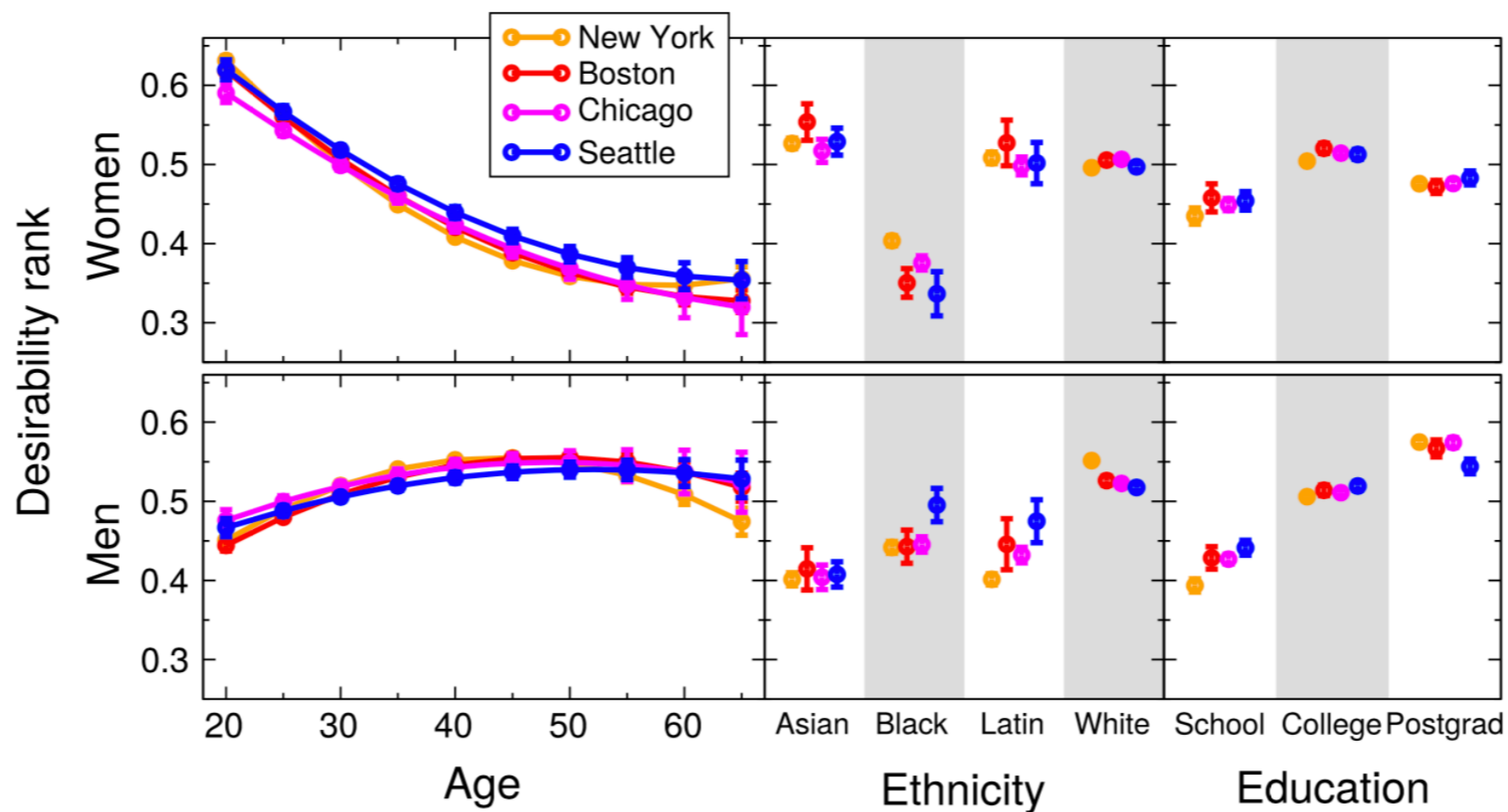
Fix: Choose the representation where we include an edge $u \rightarrow v$ if u initially contacted v , and, optionally, an edge $v \rightarrow u$ if v replied to u



Women reply <20% of the time in the data

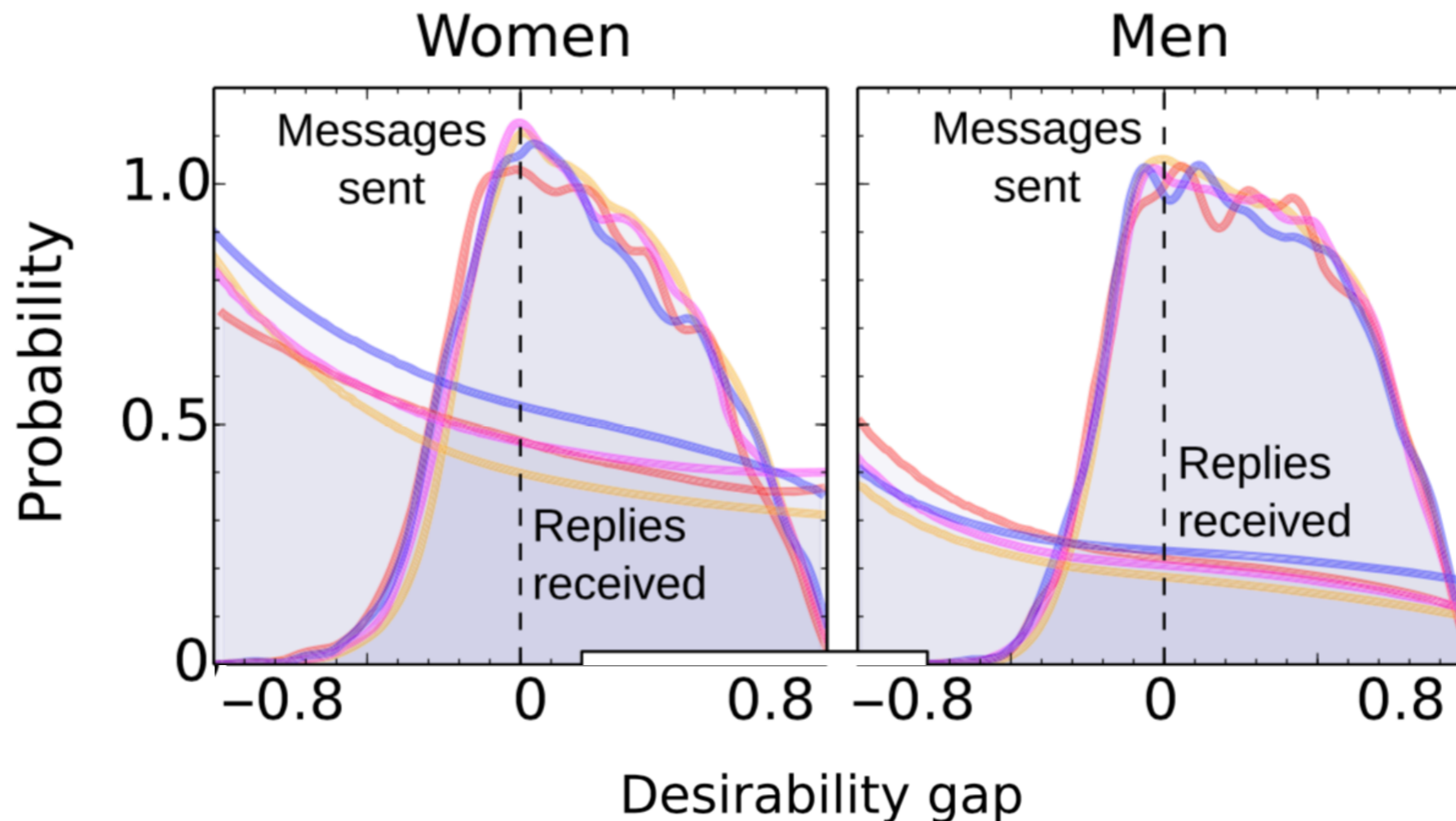
PageRank Application: Online Dating

Desirability (rank) as it varies with age, ethnicity, education



PageRank Application: Online Dating

There is a **desirability gap**: Both women and men tend to contact others who are ranked somewhat—but not excessively—higher than themselves

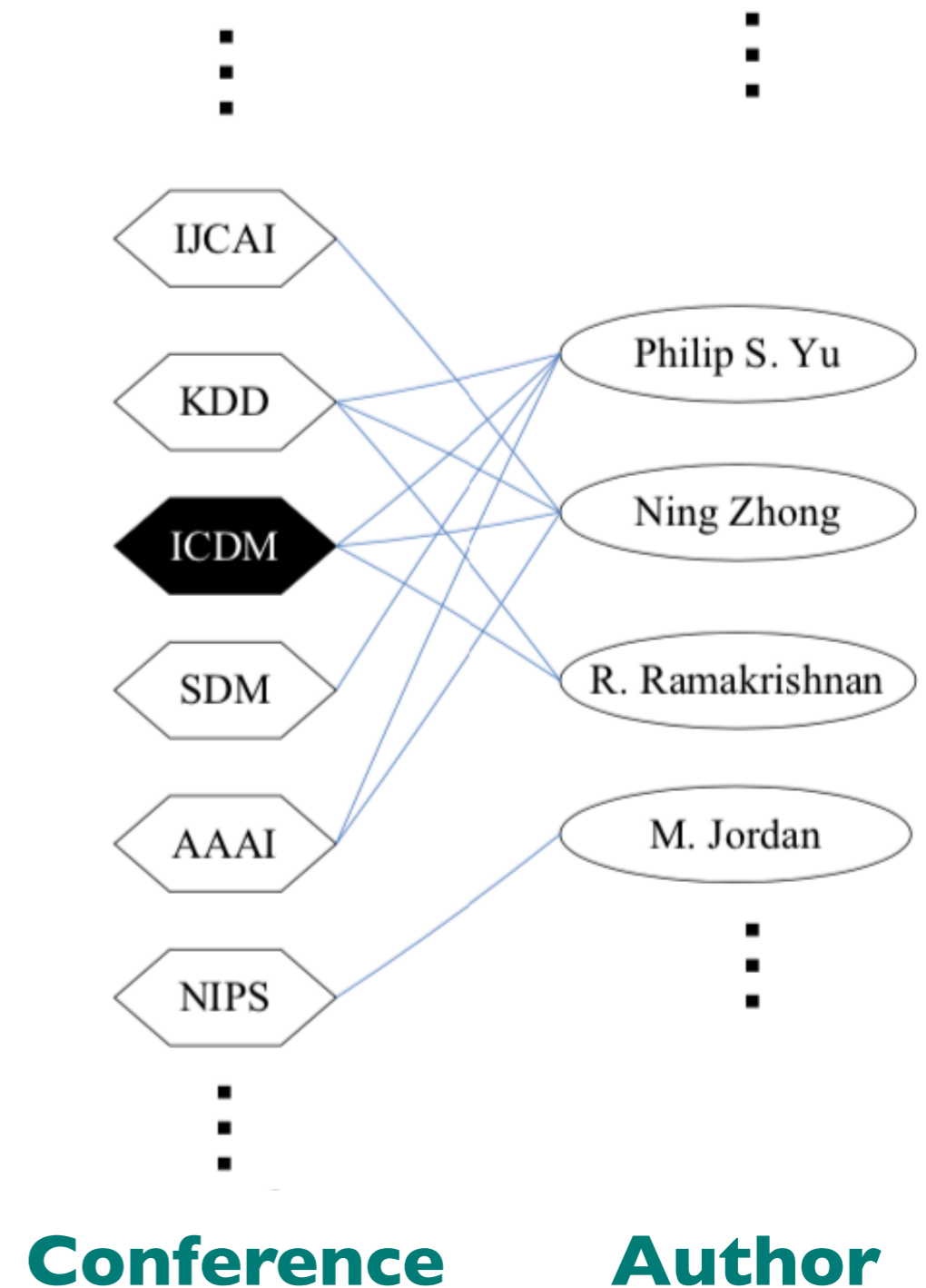


PageRank Application: Graph Search

Given: Conferences-to-authors graph

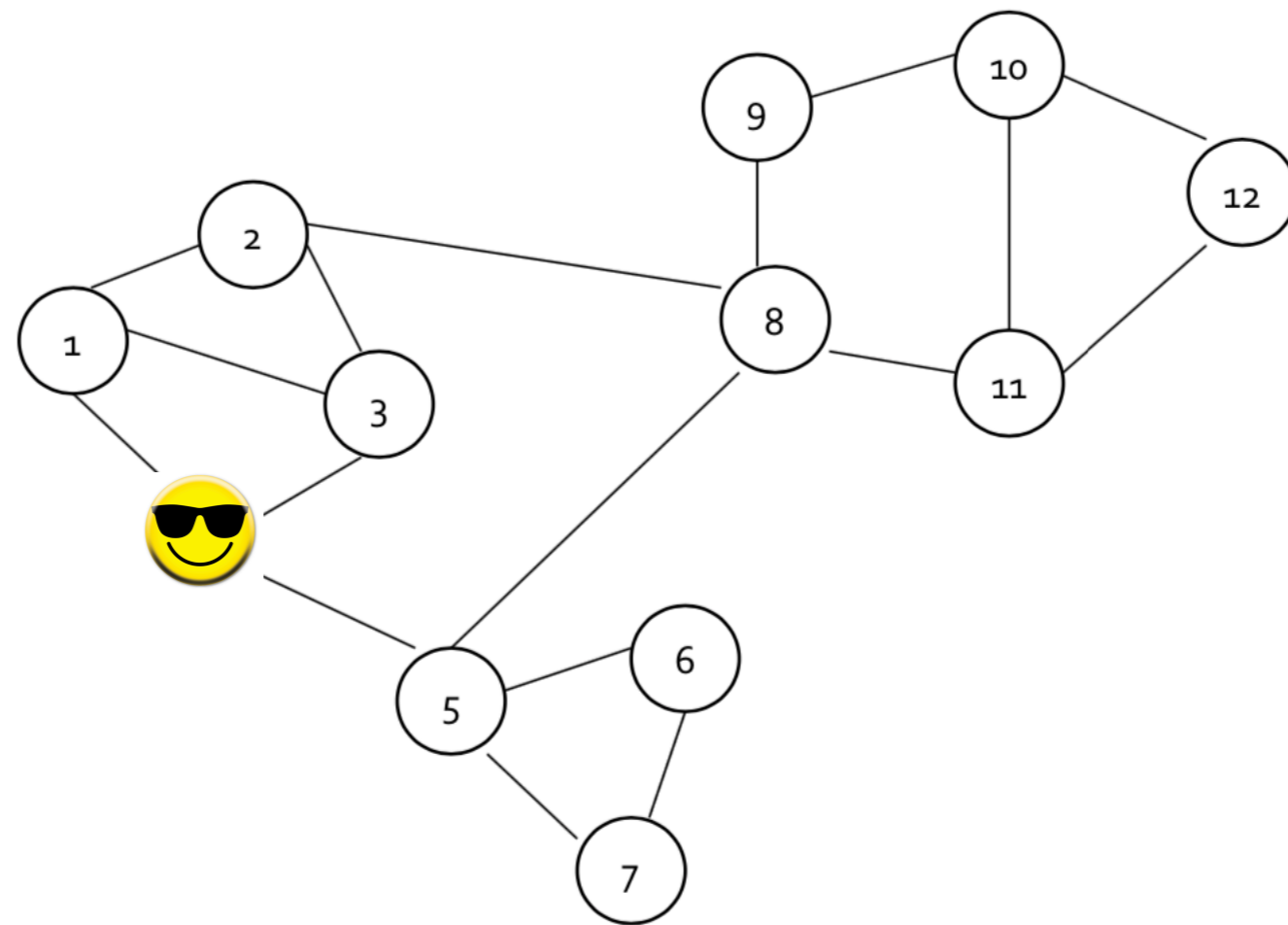
Goal: Proximity on graphs

Q: What is the most related conference to ICDM?



International Joint Conferences on Artificial Intelligence (IJCAI)
Knowledge Discovery and Data Mining (KDD)
International Conference on Data Mining (ICDM)
SIAM International Conference on Data Mining (SDM)
Association for the Advancement of Artificial Intelligence (AAAI)
Conference on Neural Information Processing Systems (NeurIPS/NIPS)

Random Walk With Restarts



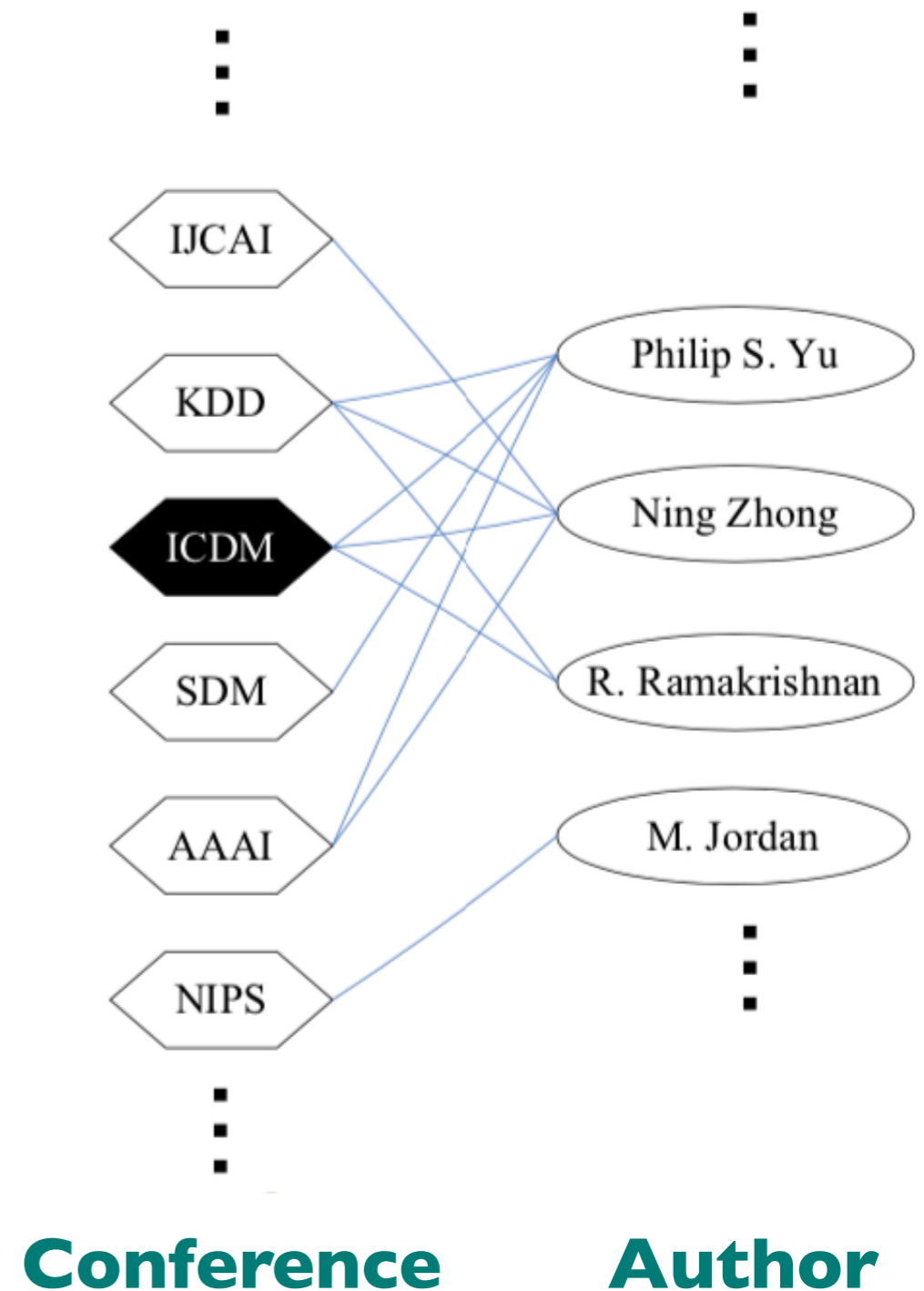
Personalized PageRank

Goal: Evaluate pages not just by popularity or global importance, but by how close they are to a given topic

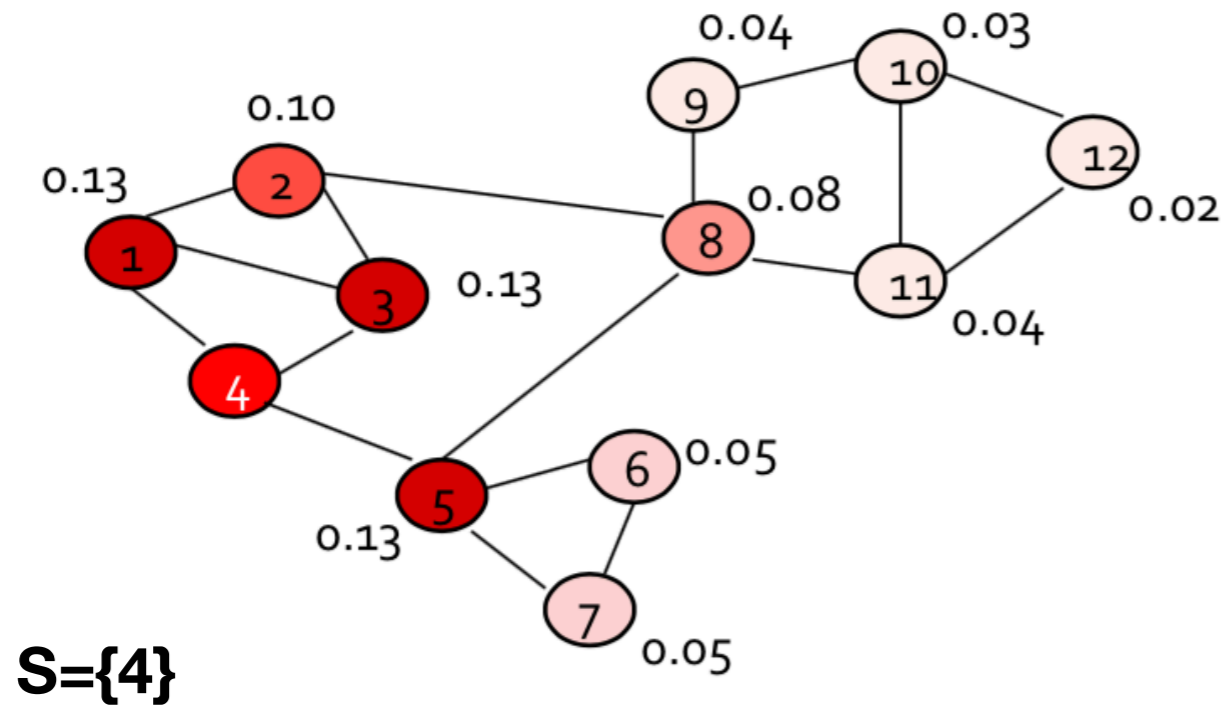
Solution: change teleportation vector!

Teleporting can go to:

- Any page with equal prob. (normal PageRank)
- A topic-specific set of “relevant” pages
- A single page/node (random walk with restarts)



Personalized PageRank

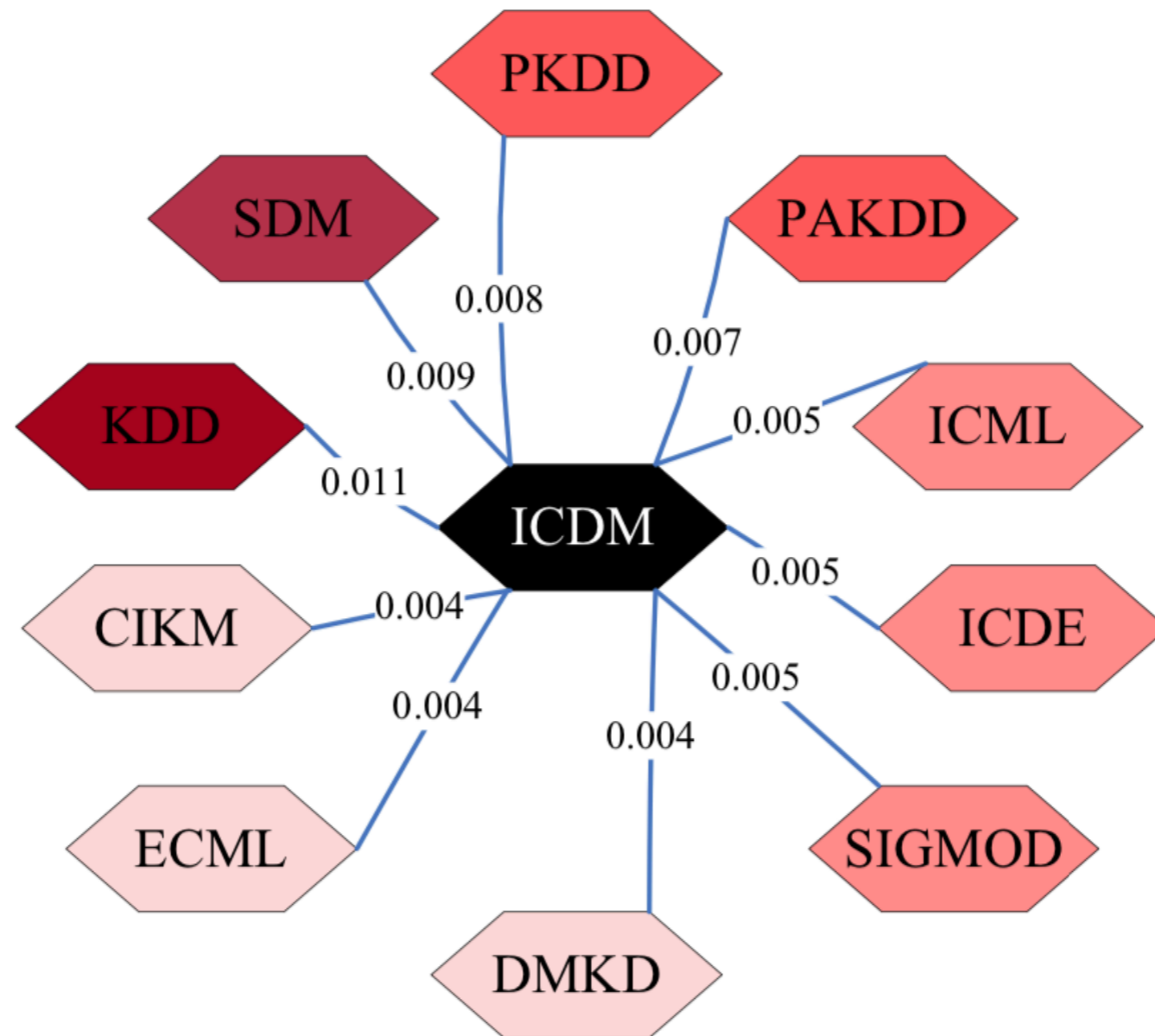


	Node 4
Node 1	0.13
Node 2	0.10
Node 3	0.13
Node 4	/
Node 5	0.13
Node 6	0.05
Node 7	0.05
Node 8	0.08
Node 9	0.04
Node 10	0.03
Node 11	0.04
Node 12	0.02

Final Personalized PageRank scores

Nearby nodes have higher scores (red)

Personalized PageRank

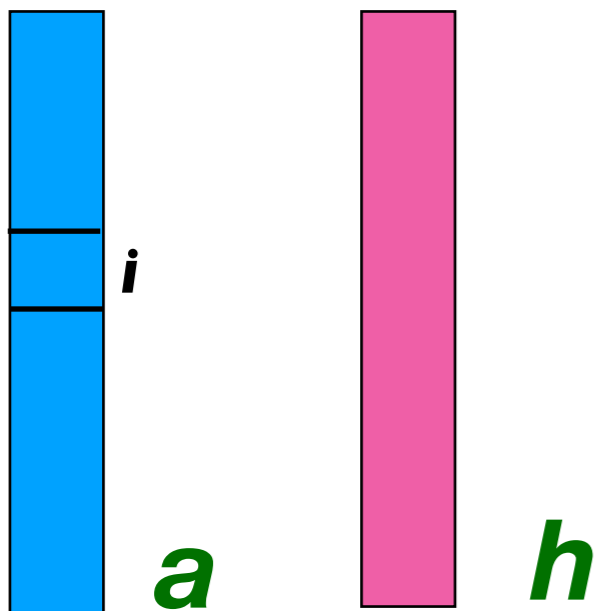
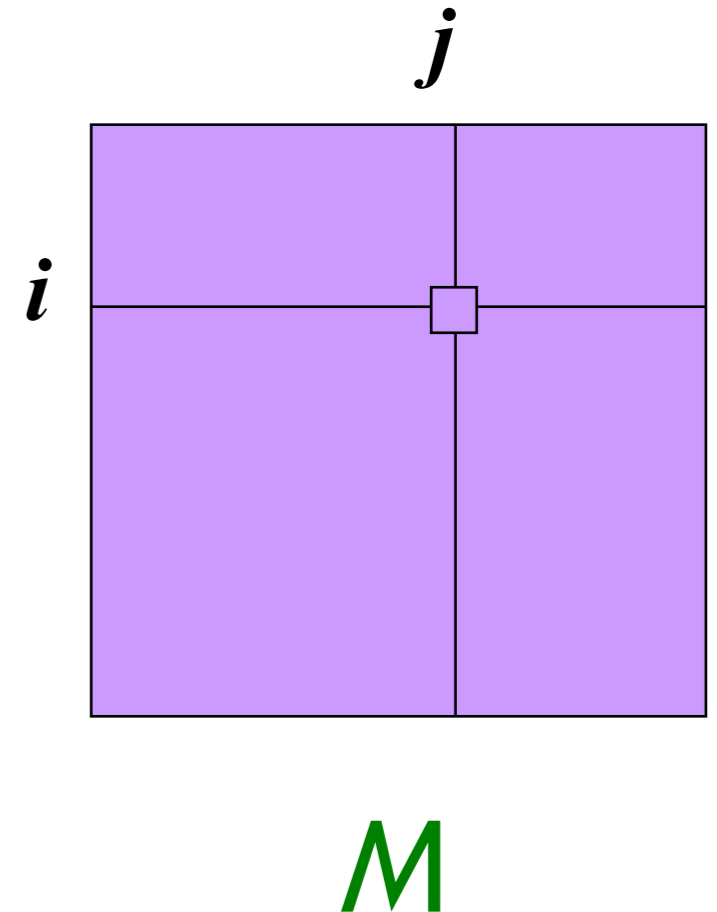


Spectral Analysis

Link Analysis: Spectral Analysis

Recall that we can represent graphs as *adjacency matrices*

$$M_{ij} = 1 \text{ if } i \text{ links to } j$$



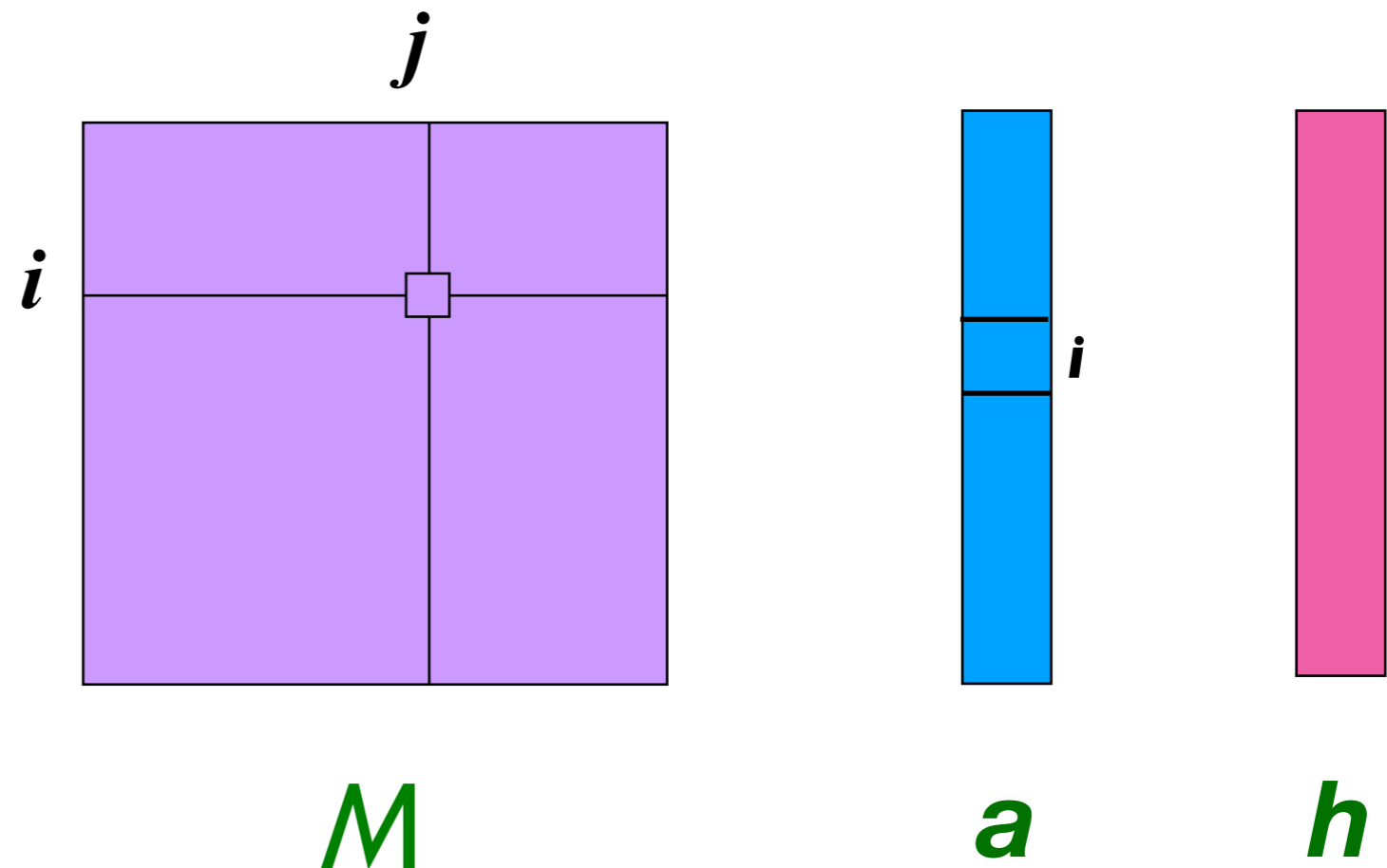
Since hub and authority scores are lists of numbers, we can represent them as vectors h and a

Link Analysis: Spectral Analysis

Q: Using M , h , and a , can you express the hub and authority update rules as matrix operations?

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

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

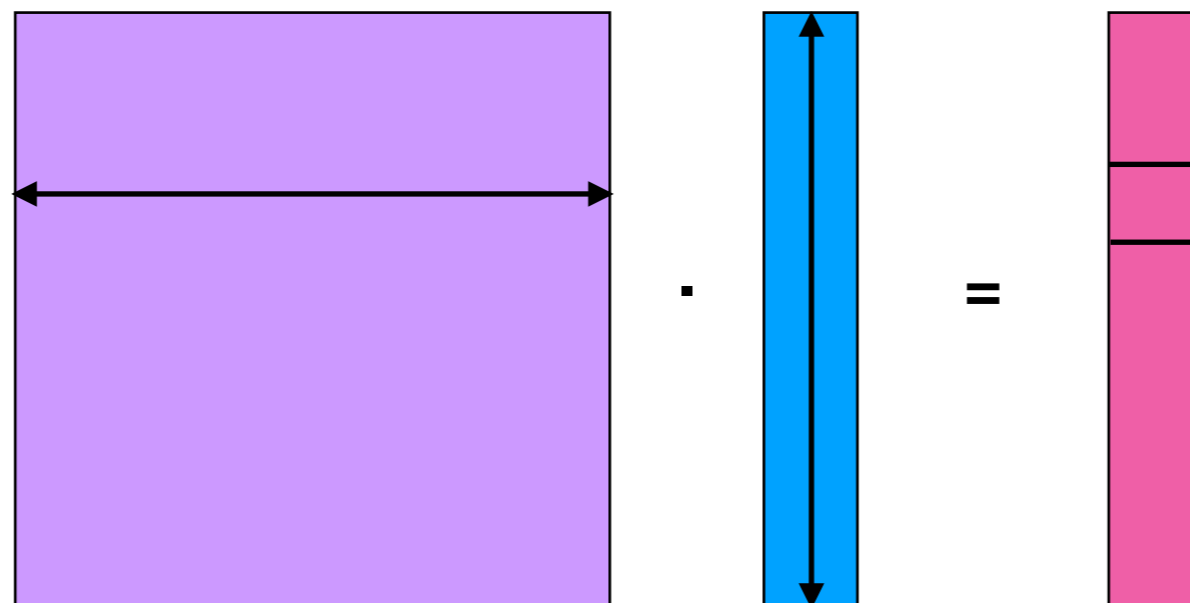


Update Rules as Matrix-Vector Multiplication

Recall Hub Update Rule:

$$h_i \leftarrow M_{i1}a_1 + M_{i2}a_2 + \dots + M_{in}a_n$$

This **corresponds exactly** to the simple matrix-vector multiplication $h \leftarrow Ma$



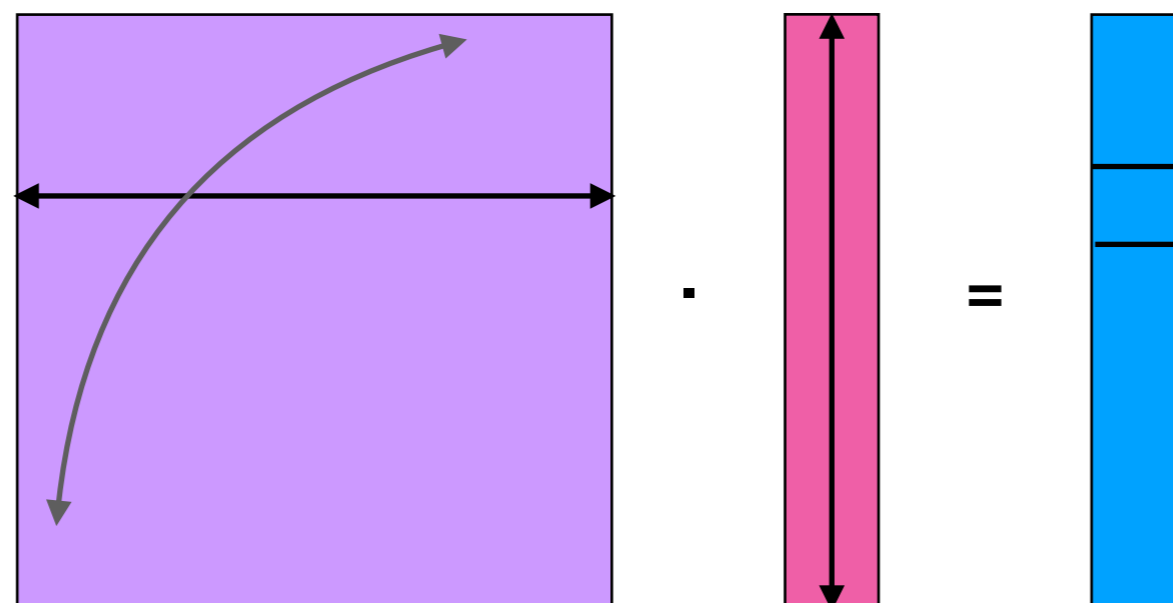
Update Rules as Matrix-Vector Multiplication

Authority update rule is similar

$$a_i \leftarrow M_{1i}h_1 + M_{2i}h_2 + \dots + M_{ni}h_n$$

This corresponds exactly to the simple matrix-vector multiplication $a \leftarrow M^T h$

Transpose the matrix!



Unwinding k steps of Updates

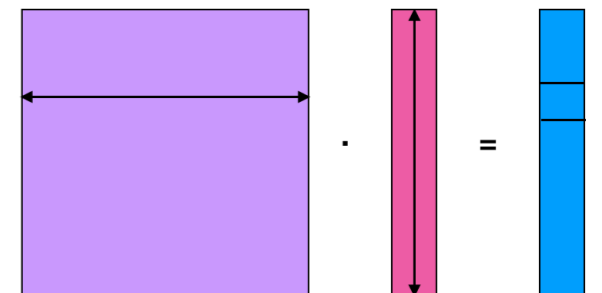
Let $a^{<j>}$ be the j -th authority vector, and similarly for $h^{<j>}$

$$a^{<1>} = M^T h^{<0>}$$

$$h^{<1>} = M a^{<1>} = M M^T h^{<0>}$$

Generally: $h^{<k>} = (M M^T)^k h^{<0>}$

Similarly: $a^{<k>} = (M^T M)^{k-1} M^T h^{<0>}$



Convergence

Recall your eigenvectors and eigenvalues:

$$Av = \lambda v$$

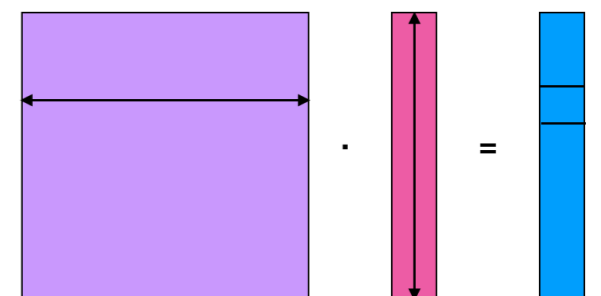
v is an eigenvector of A , with corresponding eigenvalue λ

At convergence, performing additional hub-authority steps won't change anything

Thus Hubs and Authorities converges to the leading eigenvector of MM^T and M^TM !

$$(MM^T)h^{(*)} = c \cdot h^{(*)}$$

eigenvector eigenvalue



(Full details in the reading)

Convergence

Why does it converge?

Fact: Any n-by-n **symmetric** matrix A has a **set of n eigenvectors that form a basis of \mathbb{R}^n** (i.e. they are mutually orthogonal and are all unit vectors).

Call them $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ with corresponding eigenvalues c_1, c_2, \dots, c_n (such that $|c_1| \geq |c_2| \geq \dots \geq |c_n|$, and assume $|c_1| > |c_2|$)

Easy Fact: MM^T is symmetric.

$$\text{Recall: } (AB)^T = B^T A^T$$

$$(MM^T)^T = (M^T)^T \cdot M^T = MM^T$$

MM^T is its own transpose (i.e. symmetric)

Convergence

Why does it converge?

Fact: Any n-by-n symmetric matrix A has a set of n eigenvectors that form a basis of \mathbb{R}^n (i.e. they are mutually orthogonal and are all unit vectors).

Call them z_1, z_2, \dots, z_n with corresponding eigenvalues c_1, c_2, \dots, c_n (such that $|c_1| \geq |c_2| \geq \dots \geq |c_n|$, and assume $|c_1| > |c_2|$)

Fact: MM^T is symmetric:

Now, think about any matrix-vector product with the symmetric matrix MM^T :

$$\begin{aligned}(MM^T)x &= (MM^T)(p_1 z_1 + p_2 z_2 + \dots + p_n z_n) && \text{Since the } z\text{'s form a basis for } \mathbb{R}^n \\ &= p_1 MM^T z_1 + p_2 MM^T z_2 + \dots + p_n MM^T z_n \\ &= p_1 c_1 z_1 + p_2 c_2 z_2 + \dots + p_n c_n z_n && \text{Since the } z\text{'s are eigenvectors of } MM^T\end{aligned}$$

Convergence

Rewrite Hub-Authority computation this way:

$$(MM^T)^k x = c_1^k p_1 z_1 + c_2^k p_2 z_2 + \dots + c_n^k p_n z_n$$

Every multiplication by MM^T adds a c_i to the i -th term

$$h^{(k)} = (MM^T)^k h^{(0)} = c_1^k q_1 z_1 + c_2^k q_2 z_2 + \dots + c_n^k q_n z_n$$

Rewriting $h^{(0)}$ as $\sum_i q_i z_i$

$$\frac{h^{(k)}}{c_1^k} = q_1 z_1 + \left(\frac{c_2}{c_1}\right)^k q_2 z_2 + \dots + \left(\frac{c_n}{c_1}\right)^k q_n z_n$$

Dividing both sides by c_1

$$\lim_{k \rightarrow \infty} \frac{h^{(k)}}{c_1^k} \rightarrow q_1 z_1$$

Since $|c_1| \geq |c_2| \geq \dots \geq |c_n|$, and $|c_1| > |c_2|$

We're done! Hub update converges to leading eigenvector of MM^T

Analogous analysis shows authority update converges to leading eigenvector of $M^T M$

PageRank Spectral Analysis

Recall the Basic PageRank Update Rule:

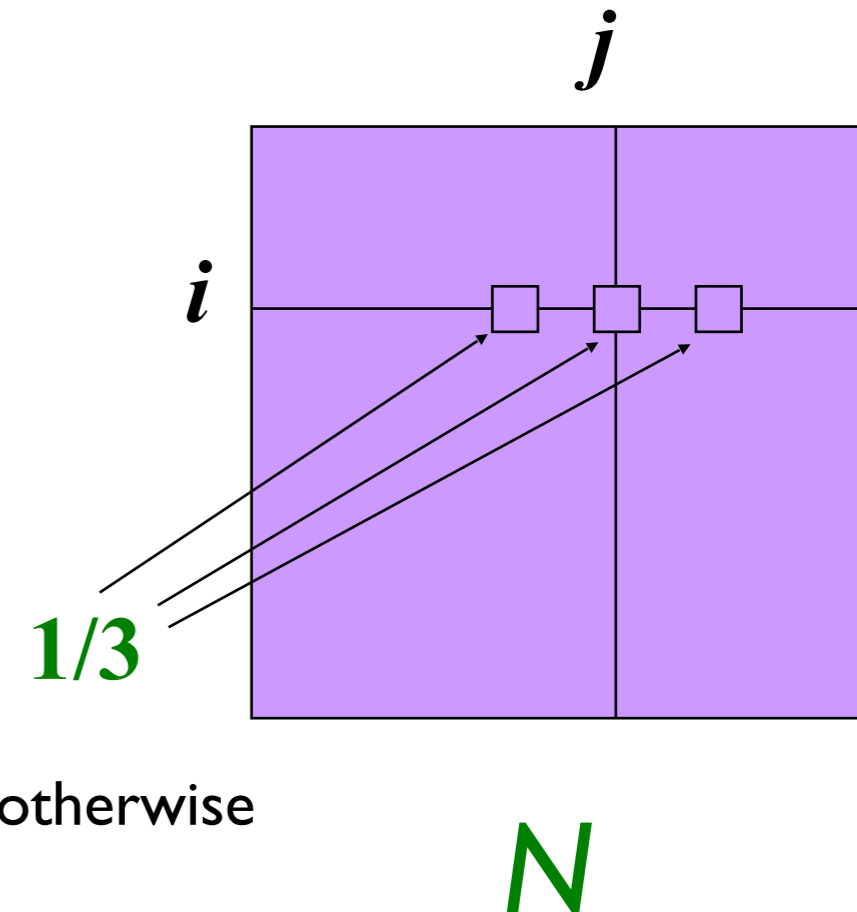
$$r_j^{\langle k+1 \rangle} = \sum_{i \rightarrow j} \frac{r_i^{\langle k \rangle}}{d_i}$$

Where $r_i^{\langle k \rangle}$ is the k-th update of i's PageRank, and node i has d_i out-links

PageRank Spectral Analysis

Recall the Basic PageRank Update Rule:

$$r_j^{(k+1)} = \sum_{i \rightarrow j} \frac{r_i^{(k)}}{d_i}$$



Define a new matrix N : $N_{ij} = \frac{1}{d_i}$ for edges $i \rightarrow j$, 0 otherwise

where page i has d_i out-links

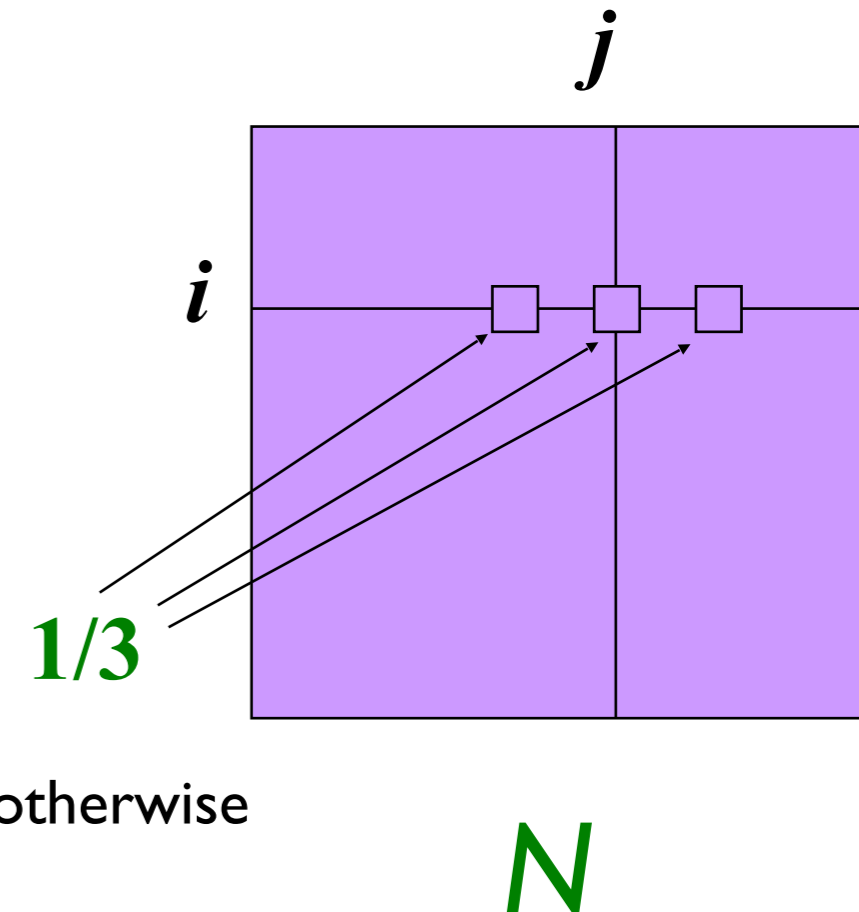
And let $r^{(k)}$ be the vector of PageRank values after k Basic PageRank Updates, where $\sum_i r_i^{(k)} = 1$

Can you write down the basic PageRank rule in terms of N and $r^{(k)}$?

PageRank Spectral Analysis

Recall the Basic PageRank Update Rule:

$$r_j^{(k+1)} = \sum_{i \rightarrow j} \frac{r_i^{(k)}}{d_i}$$



Define a new matrix N : $N_{ij} = \frac{1}{d_i}$ for edges $i \rightarrow j$, 0 otherwise

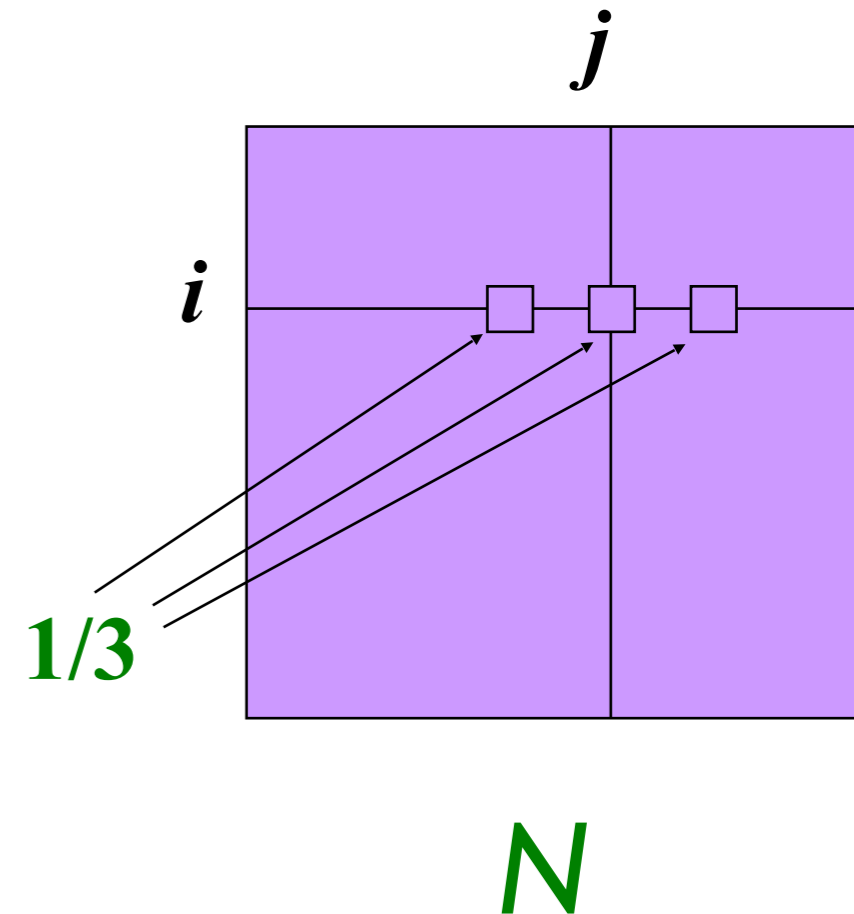
where page i has
 d_i out-links

$$r^{(k+1)} = N_{1i}r_1^{(k)} + N_{2i}r_2^{(k)} + \dots + N_{ni}r_n^{(k)}$$

$$r^{(k+1)} = N^T r^{(k)}$$

PageRank Spectral Analysis

Similarly, PageRank converges to the leading eigenvector of N^T



PageRank and HITS

PageRank and HITS are two solutions to the same problem:

What is the value of an in-link from u to v ?

In the PageRank model, the value of the link depends on the links into u

In the HITS model, it depends on the value of the other links out of u

The destinies of PageRank and HITS post-1998 were very different