



Introduction to Data Mining

Lecture #11: Link Analysis-3

U Kang
Seoul National University



Outline

- ➔ **Web Spam: Overview**
- TrustRank: Combating the Web Spam
- HITS: Hubs and Authorities



What is Web Spam?

- **Spamming:**
 - Any deliberate action to boost a web page's position in search engine results, incommensurate with page's real value
- **Spam:**
 - Web pages that are the result of spamming
- This is a very broad definition
 - **SEO** industry might disagree!
 - SEO = search engine optimization
- Approximately **10-15%** of web pages are spam



Web Search

■ Early search engines:

- ❑ Crawl the Web
- ❑ Index pages by the words they contained
- ❑ Respond to search queries (lists of words) with the pages containing those words

■ Early page ranking:

- ❑ Attempt to order pages matching a search query by “importance”
- ❑ **First search engines considered:**
 - (1) Number of times query words appeared
 - (2) Prominence of word position, e.g. title, header



First Spammers

- As people began to use search engines to find things on the Web, those with commercial interests tried to **exploit search engines** to bring people to their own site – whether they wanted to be there or not
- **Example:**
 - Shirt-seller might pretend to be about “movies”
- **Techniques for achieving high relevance/importance for a web page**



First Spammers: Term Spam

- **How do you make your page appear to be about movies?**
 - ❑ **(1)** Add the word `movie` 1,000 times to your page
 - ❑ Set text color to the background color, so only search engines would see it
 - ❑ **(2)** Or, run the query “movie” on your target search engine
 - ❑ See what page came first in the listings
 - ❑ Copy it into your page, make it “invisible”
- **These and similar techniques are term spam**



Google's Solution to Term Spam

- Believe what people say about you, rather than what you say about yourself
 - Use words in the anchor text (words that appear underlined to represent the link) and its surrounding text
- PageRank as a tool to measure the “importance” of Web pages



Why It Works?

■ Our hypothetical shirt-seller loses

- ❑ Saying he is about movies doesn't help, because others don't say he is about movies
- ❑ His page isn't very important, so it won't be ranked high for shirts or movies

■ Example:

- ❑ Shirt-seller creates 1,000 pages, each links to his page with "movie" in the anchor text
- ❑ These pages have no links in, so they get little PageRank
- ❑ So the shirt-seller can't beat truly important movie pages, like IMDB



SPAM FARMING



Google vs. Spammers: Round 2!

- Once Google became the dominant search engine, spammers began to work out ways to fool Google
- **Spam farms** were developed to concentrate PageRank on a single page
- **Link spam:**
 - Creating link structures that boost PageRank of a particular page





Link Spamming

- **Three kinds of web pages from a spammer's point of view**
 - **Inaccessible pages**
 - spammer has no control
 - **Accessible pages**
 - e.g., blog comments pages
 - spammer can post links to his pages
 - **Owned pages**
 - Completely controlled by spammer
 - May span multiple domain names



Link Farms

- **Spammer's goal:**

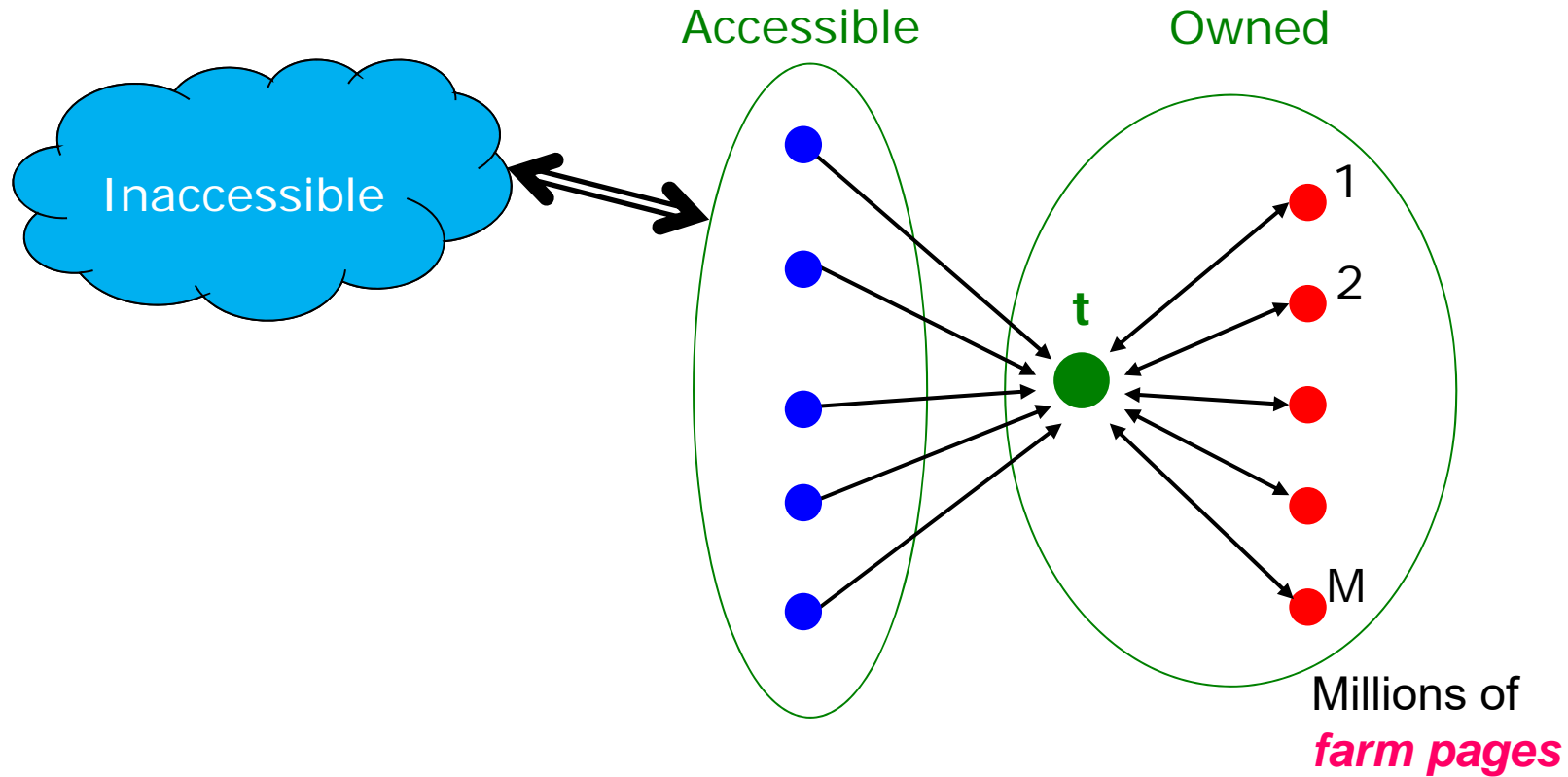
- Maximize the PageRank of target page t

- **Technique:**

- Get as many links from accessible pages as possible to target page t
- Construct “link farm” using owned pages to get PageRank multiplier effect



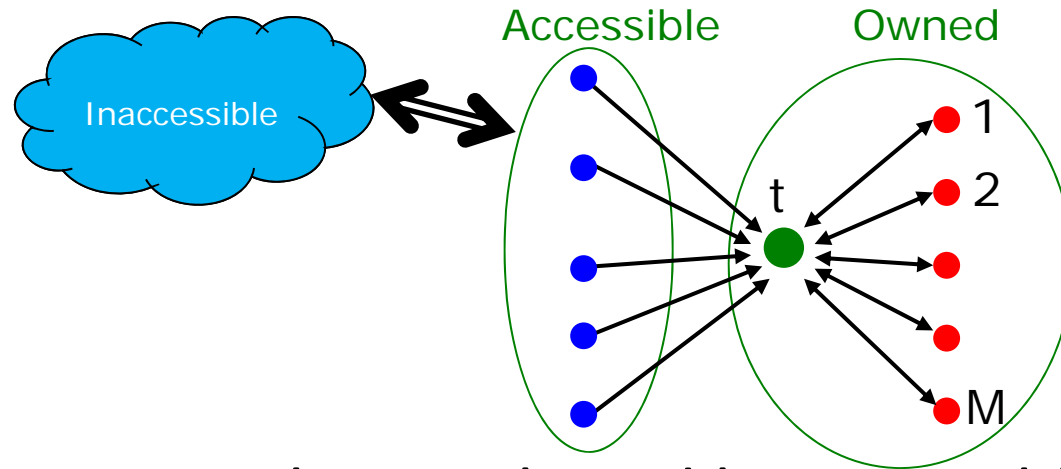
Link Farms



One of the most common and effective organizations for a link farm



Analysis



N...# pages on the web
M...# of pages spammer owns

- x : PageRank contributed by accessible pages

- y : PageRank of target page t

- Rank of each “farm” page = $\frac{\beta y}{M} + \frac{1-\beta}{N}$

- $y = x + \beta M \left[\frac{\beta y}{M} + \frac{1-\beta}{N} \right] + \frac{1-\beta}{N}$

$$= x + \beta^2 y + \frac{\beta(1-\beta)M}{N} + \frac{1-\beta}{N}$$

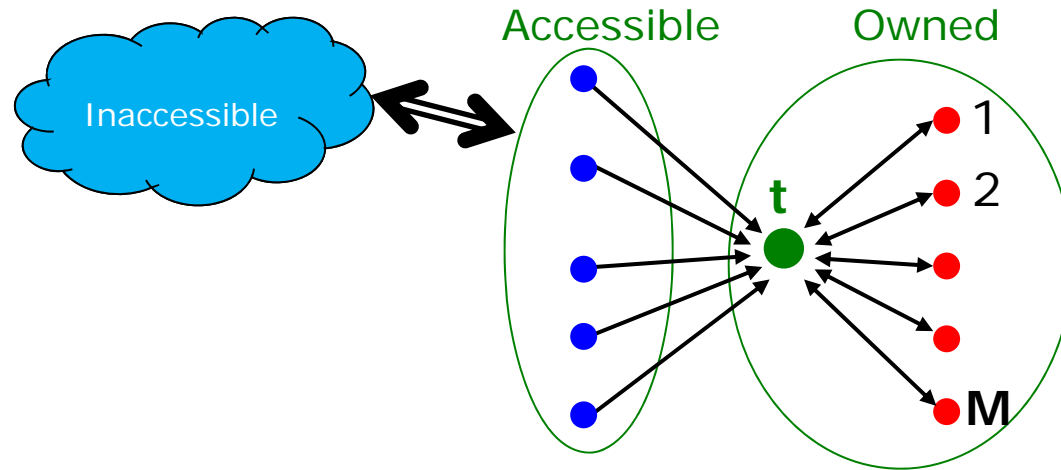
Very small; ignore
Now we solve for y

- $y = \frac{x}{1-\beta^2} + c \frac{M}{N}$ where $c = \frac{\beta}{1+\beta}$

U Kang



Analysis




N...# pages on the web
M...# of pages spammer owns

- $y = \frac{x}{1-\beta^2} + c \frac{M}{N}$ where $c = \frac{\beta}{1+\beta}$
- For $\beta = 0.85$, $1/(1-\beta^2) = 3.6$
- Multiplier effect for acquired PageRank
- By making M large, we can make y as large as we want (up to c)



Outline

- Web Spam: Overview
-  **TrustRank: Combating the Web Spam**
- HITS: Hubs and Authorities



Combating Spam

- **Combating term spam**
 - Use anchor text, and PageRank
 - Analyze text using statistical methods
 - Also useful: Detecting approximate duplicate pages
- **Combating link spam**
 - **Detection and blacklisting of structures that look like spam farms**
 - Leads to another war – hiding and detecting spam farms
 - **TrustRank** = topic-specific PageRank with a teleport set of **trusted pages**
 - **Example:** .edu domains, similar domains for non-US schools



TrustRank: Idea

- **Basic principle: Approximate isolation**
 - It is rare for a “good” page to point to a “bad” (spam) page
- Sample a set of **seed pages** from the web
- Have an **oracle (human)** to identify the good pages and the spam pages in the seed set
 - **Expensive task**, so we must make seed set as small as possible



Trust Propagation

- Call the subset of seed pages that are identified as **good** the **trusted pages**
- Perform a topic-sensitive PageRank with **teleport set = trusted pages**
 - **Propagate trust through links:**
 - Each page gets a trust value between **0** and **1**
- **Solution 1: Use a threshold value and mark all pages below the trust threshold as spam**



Why is it a good idea?

■ Trust attenuation:

- The degree of trust conferred by a trusted page decreases with the distance in the graph

■ Trust splitting:

- The larger the number of out-links from a page, the less trust the page author gives to each out-link
- Trust is **split** across out-links



Picking the Seed Set

- **Two conflicting considerations:**
 - Human has to inspect each seed page, so seed set must be as small as possible
 - Must ensure every **good page** gets adequate trust rank, so need to make all good pages reachable from seed set by short paths



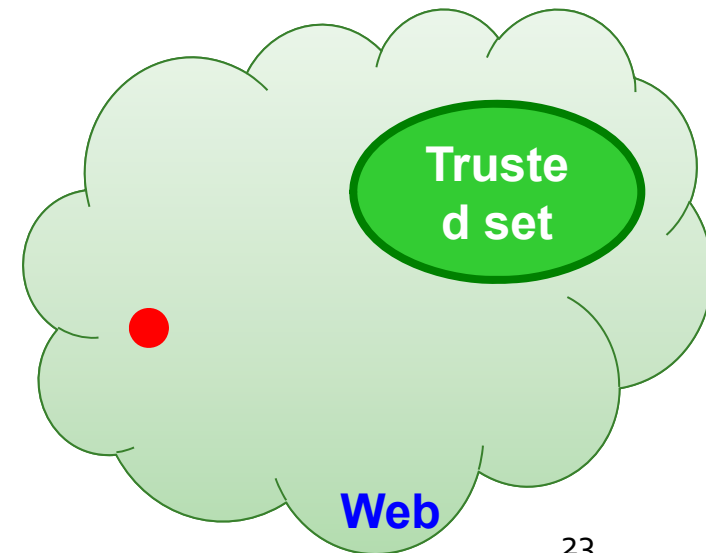
Approaches to Picking Seed Set

- Suppose we want to pick a seed set of k pages
- **How to do that?**
- **(1) PageRank:**
 - Pick the top k pages by PageRank
 - Main idea: you can't get a bad page's rank really high
- **(2) Use trusted domains** whose membership is controlled, like .edu, .mil, .gov



Spam Mass

- In the **TrustRank** model, we start with good pages and propagate trust
- **Complementary view:**
What fraction of a page's PageRank comes from **spam** pages?
- In practice, we don't know all the spam pages, so we need to estimate





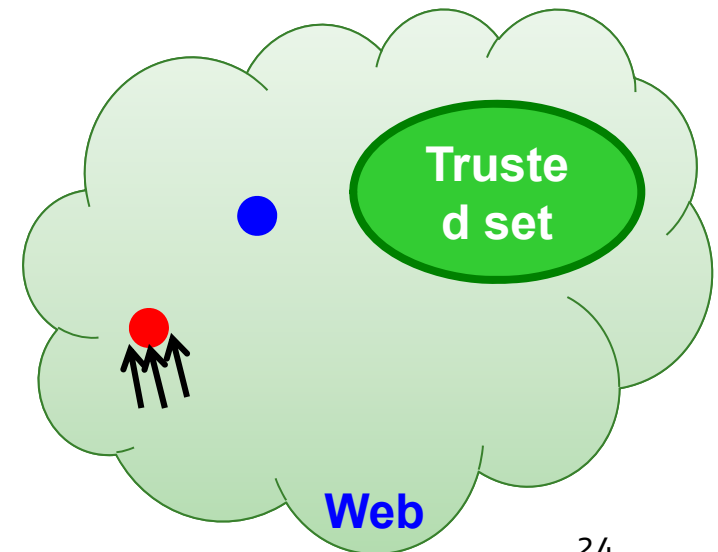
Spam Mass Estimation

Solution:

- r_p = PageRank of page p
- r_p^+ = PageRank of p with teleport into **trusted** pages only
- **Then:** What fraction of a page's PageRank comes from **spam** pages?

$$r_p^- = r_p - r_p^+$$

- **Spam mass of p** = $\frac{r_p^-}{r_p}$
 - Pages with high spam mass are spam.





Outline

- Web Spam: Overview
- TrustRank: Combating the Web Spam
- HITS: Hubs and Authorities**



Hubs and Authorities

- **HITS (Hypertext-Induced Topic Selection)**

- Is a measure of importance of pages or documents, similar to PageRank
- Proposed at around same time as PageRank ('98)

- **Goal:** Say we want to find good newspapers

- Don't just find newspapers. Find “experts” – people who link in a coordinated way to good newspapers

- **Idea: Links as votes**

- **Page is more important if it has more links**
 - In-coming links? Out-going links?



Finding newspapers

■ Hubs and Authorities

Each page has 2 scores:

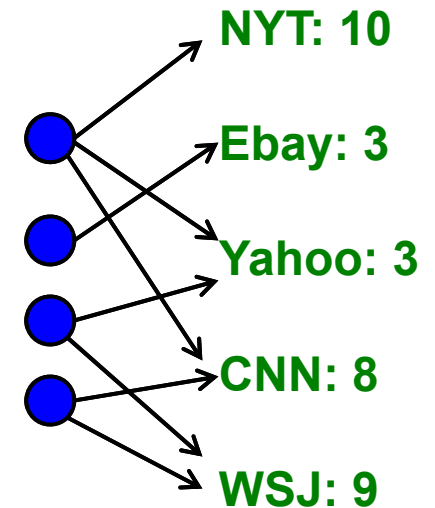
□ Quality as an expert (**hub**):

- Total sum of votes of authorities it points to

□ Quality as a content (**authority**):

- Total sum of votes coming from experts

■ Principle of repeated improvement

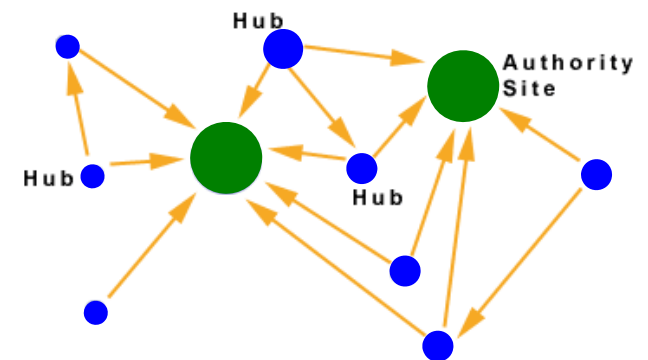




Hubs and Authorities

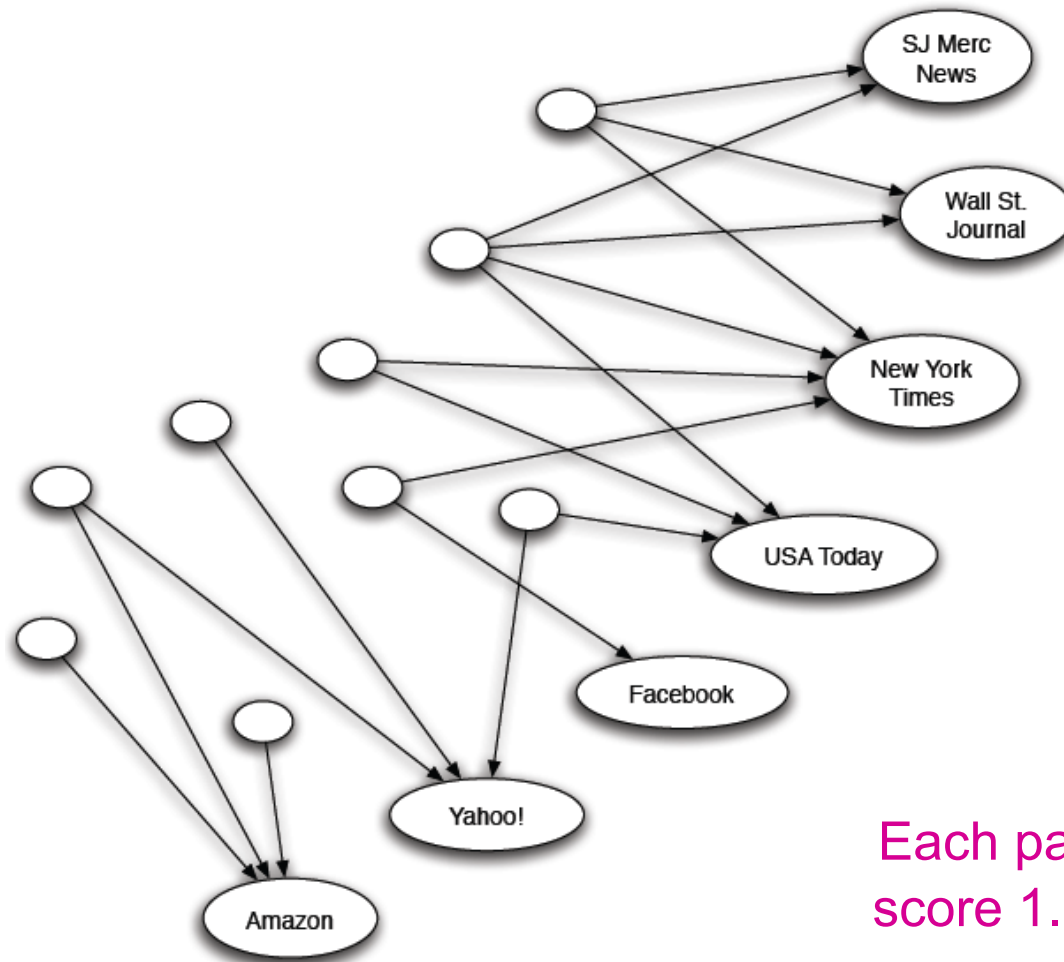
Interesting pages fall into two classes:

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 US auto manufacturers





Counting in-links: Authority

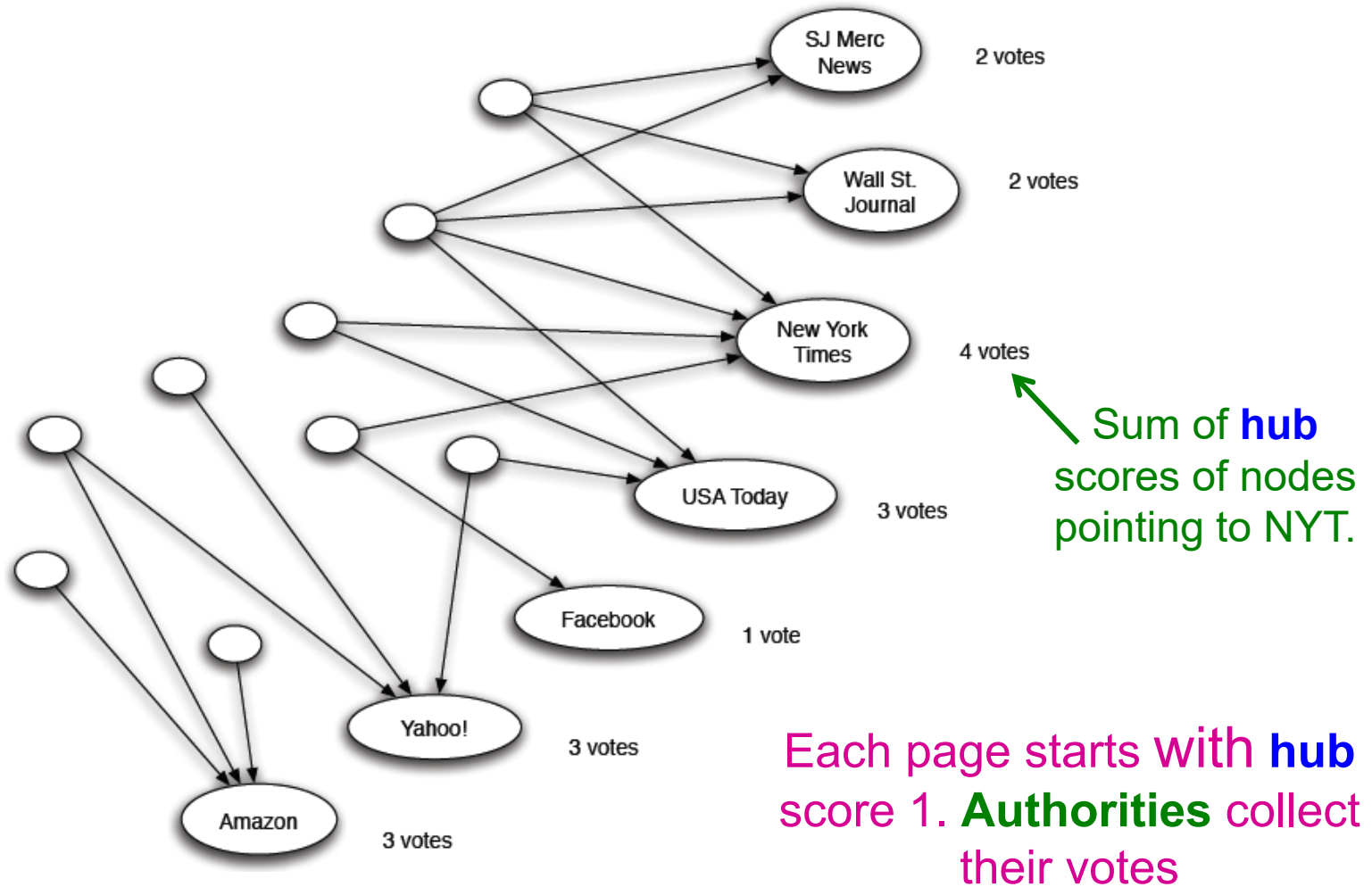


Each page starts with **hub** score 1. **Authorities** collect their votes

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)



Counting in-links: Authority

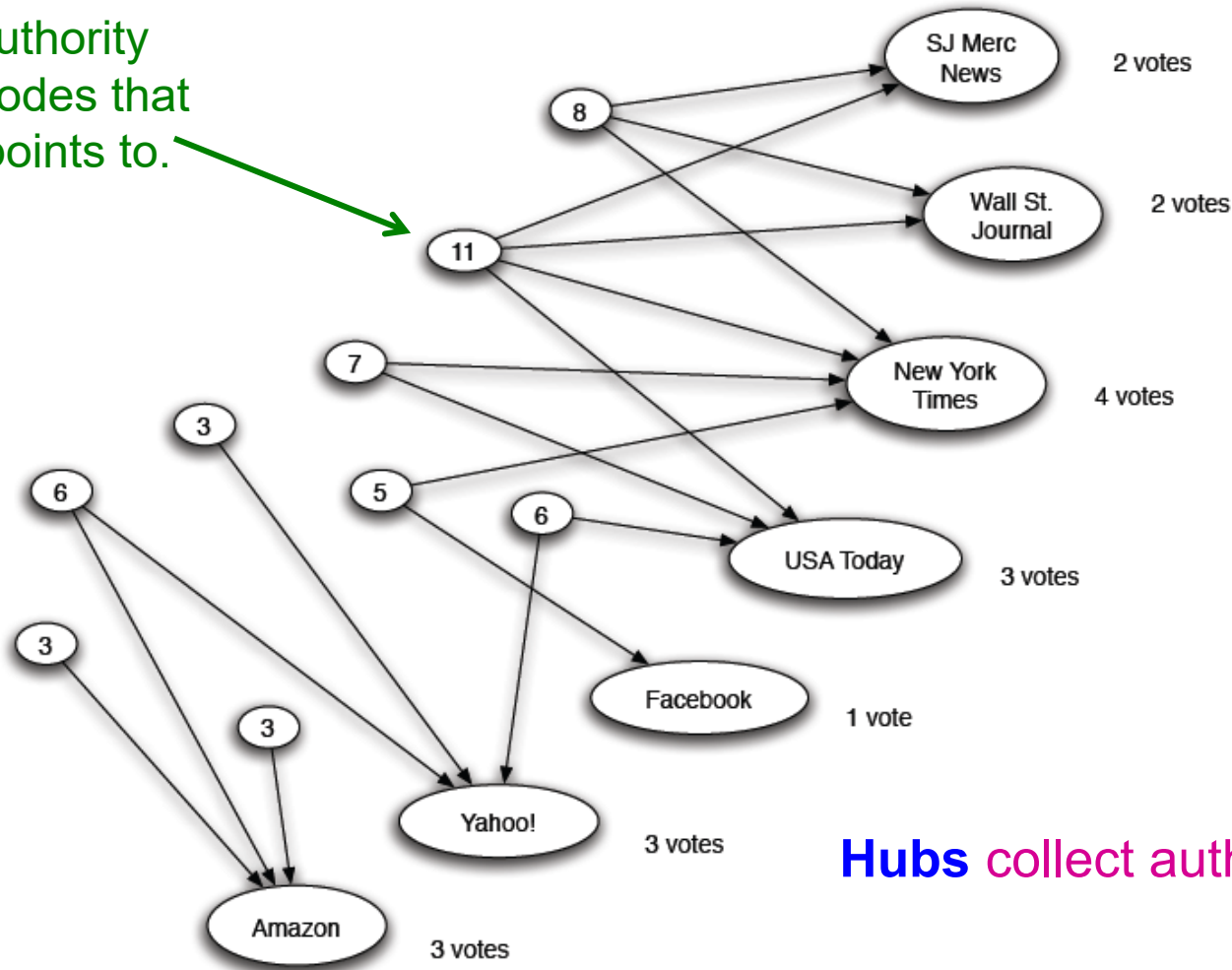


(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)



Expert Quality: Hub

Sum of authority scores of nodes that the node points to.

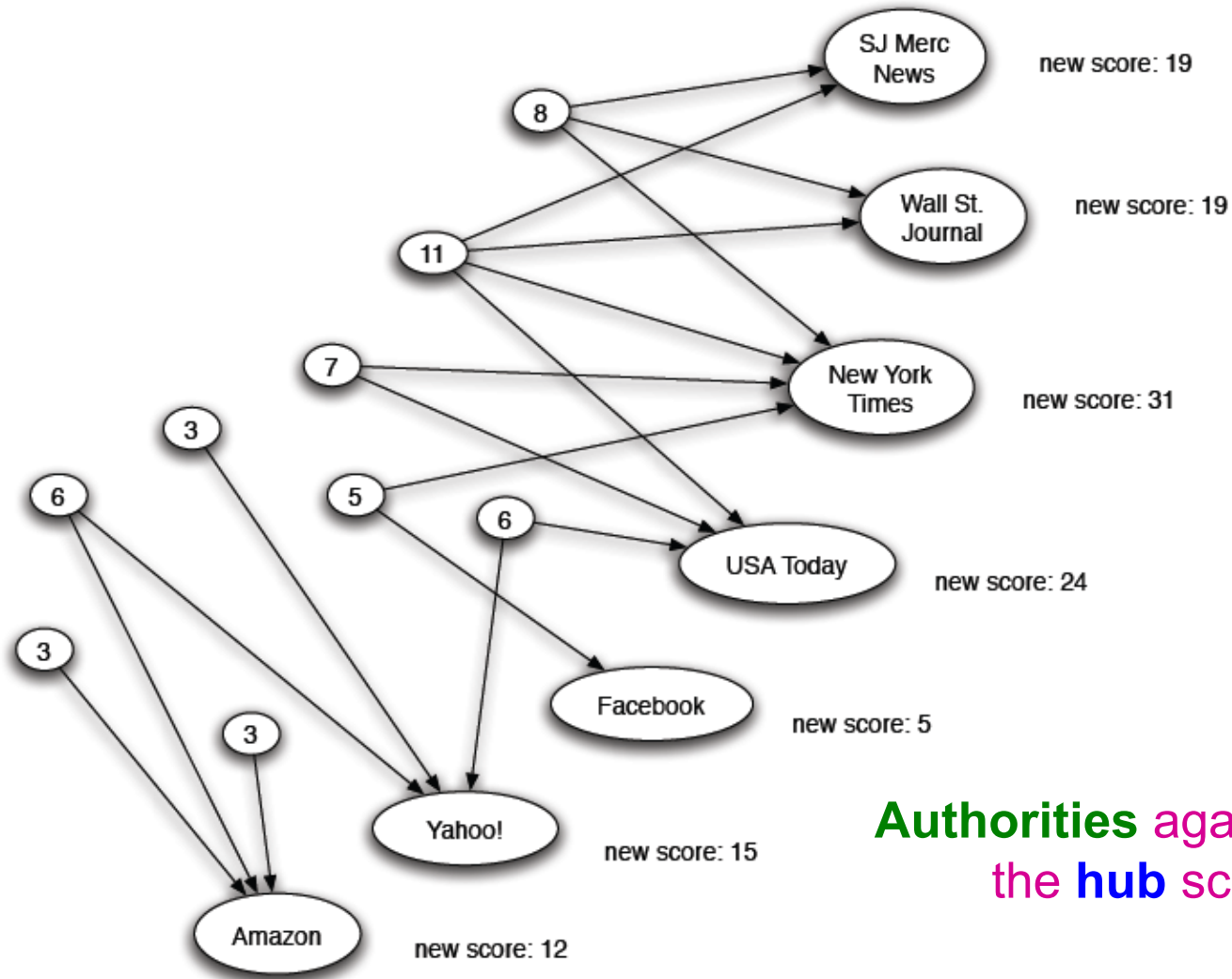


Hubs collect authority scores

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)



Reweighting



Authorities again collect the **hub** scores

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)



Mutually Recursive Definition

- A good hub links to many good authorities
- A good authority is linked from many good hubs
- Model using two scores for each node:
 - Hub score and Authority score
 - Represented as vectors h and a



HITS

Then:

$$a_i = h_k + h_l + h_m$$

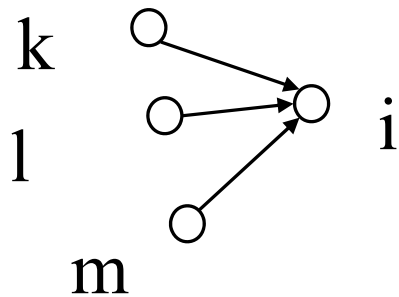
that is

$a_i = \text{Sum } (h_j)$ over all j that edge (j,i) exists

or

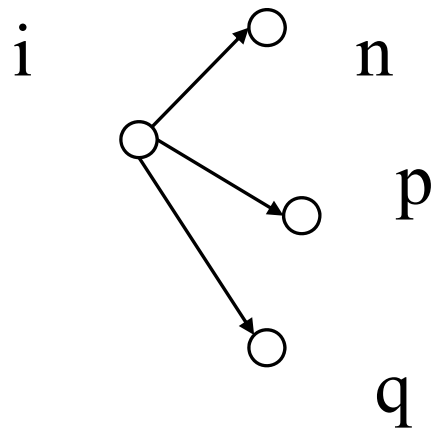
$$\mathbf{a} = \mathbf{A}^T \mathbf{h}$$

Where A is the adjacency matrix
 (i,j) is 1 if the edge from i to j exists





HITS



symmetrically, for the 'hubness' :

$$h_i = a_n + a_p + a_q$$

that is

$h_i = \text{Sum } (q_j)$ over all j that edge (i,j) exists

or

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



HITS

In conclusion, we want vectors \mathbf{h} and \mathbf{a} such that:

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

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

$$\begin{bmatrix} \\ \\ \end{bmatrix} = \begin{bmatrix} & \\ & \\ & \end{bmatrix} \begin{bmatrix} \\ \end{bmatrix}$$



HITS

In short, the solutions to

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

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

are the largest eigenvectors of $\mathbf{A} \mathbf{A}^T$ and $\mathbf{A}^T \mathbf{A}$.

Starting from random \mathbf{a}' and iterating, we'll eventually converge



HITS

Convergence: why?

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

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

$$\mathbf{h} = \mathbf{A} \mathbf{A}^T \mathbf{h} = (\mathbf{A} \mathbf{A}^T) \mathbf{h}$$

$$\mathbf{a} = \mathbf{A}^T \mathbf{A} \mathbf{a} = (\mathbf{A}^T \mathbf{A}) \mathbf{a}$$

That is, after many iterations,

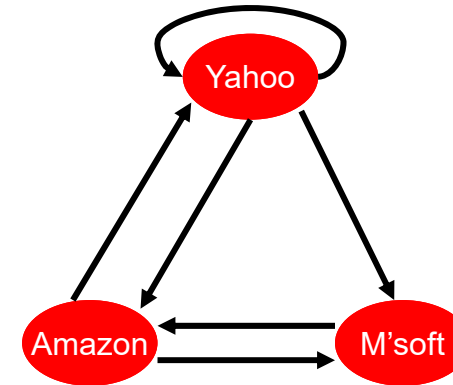
- \mathbf{h} converges to the largest eigenvector of $\mathbf{A} \mathbf{A}^T$
- \mathbf{a} converges to the largest eigenvector of $\mathbf{A}^T \mathbf{A}$



Example of HITS

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$



$h(\text{yahoo})$	$=$.58	.80	.80	.79788
$h(\text{amazon})$	$=$.58	.53	.53	.57577
$h(\text{m'soft})$	$=$.58	.27	.27	.23211
$a(\text{yahoo})$	$=$.58	.58	.62	.62628
$a(\text{amazon})$	$=$.58	.58	.49	.49459
$a(\text{m'soft})$	$=$.58	.58	.62	.62628



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



Questions?