Networks Link Analysis: Hubs and Authorities

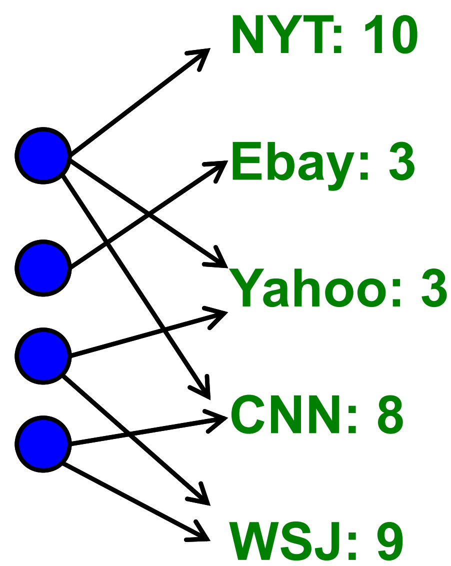
Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

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 pointed to
  - **Quality as a content (authority):**
    - Total sum of votes coming from experts



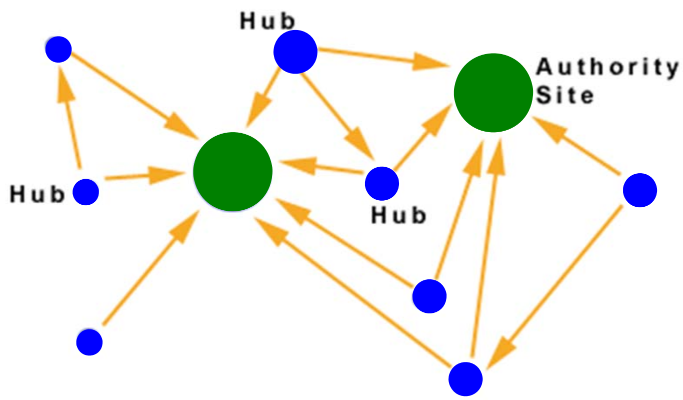
- **Principle of repeated improvement**

Hubs and Authorities

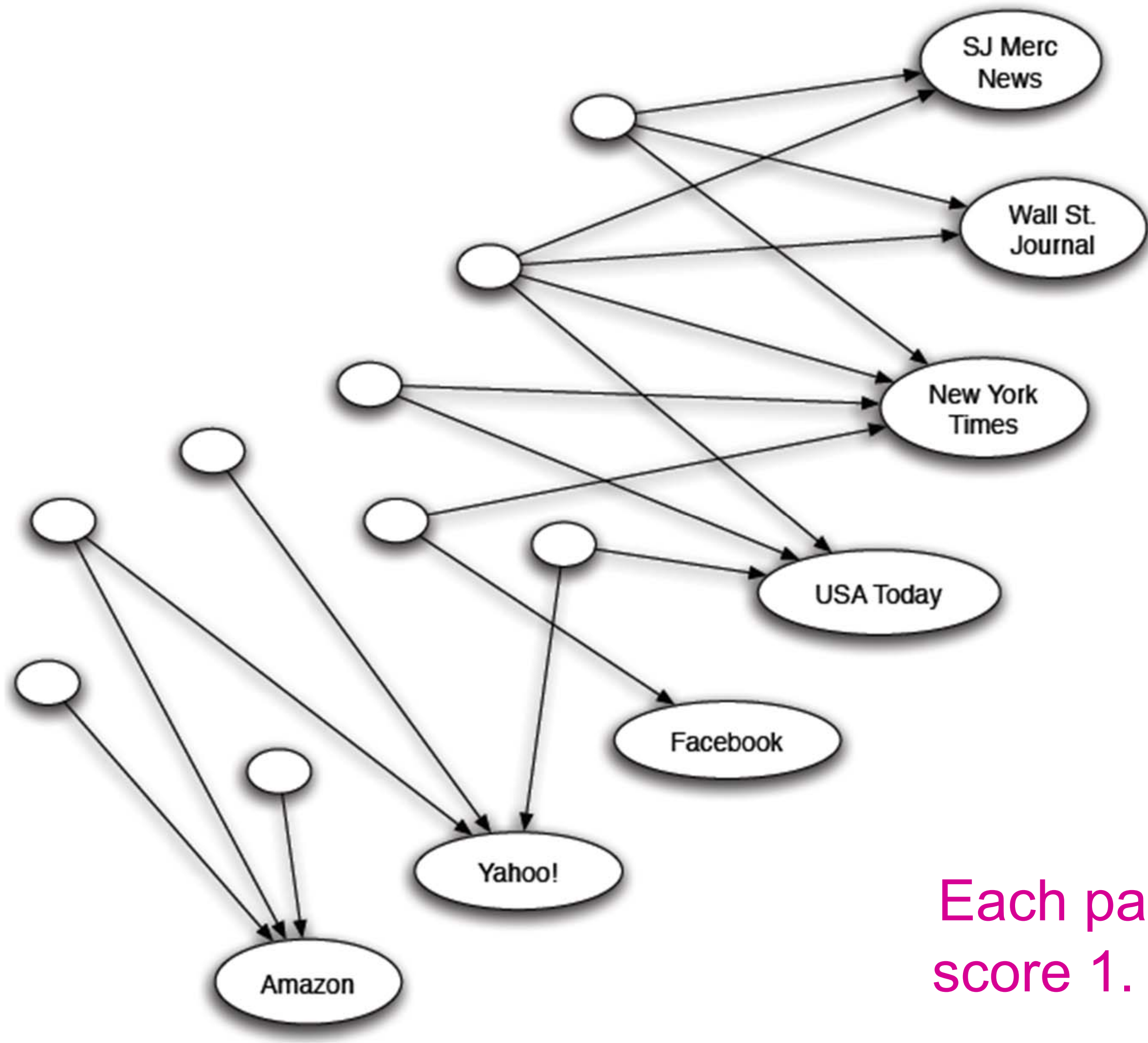
Interesting pages fall into two classes:

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



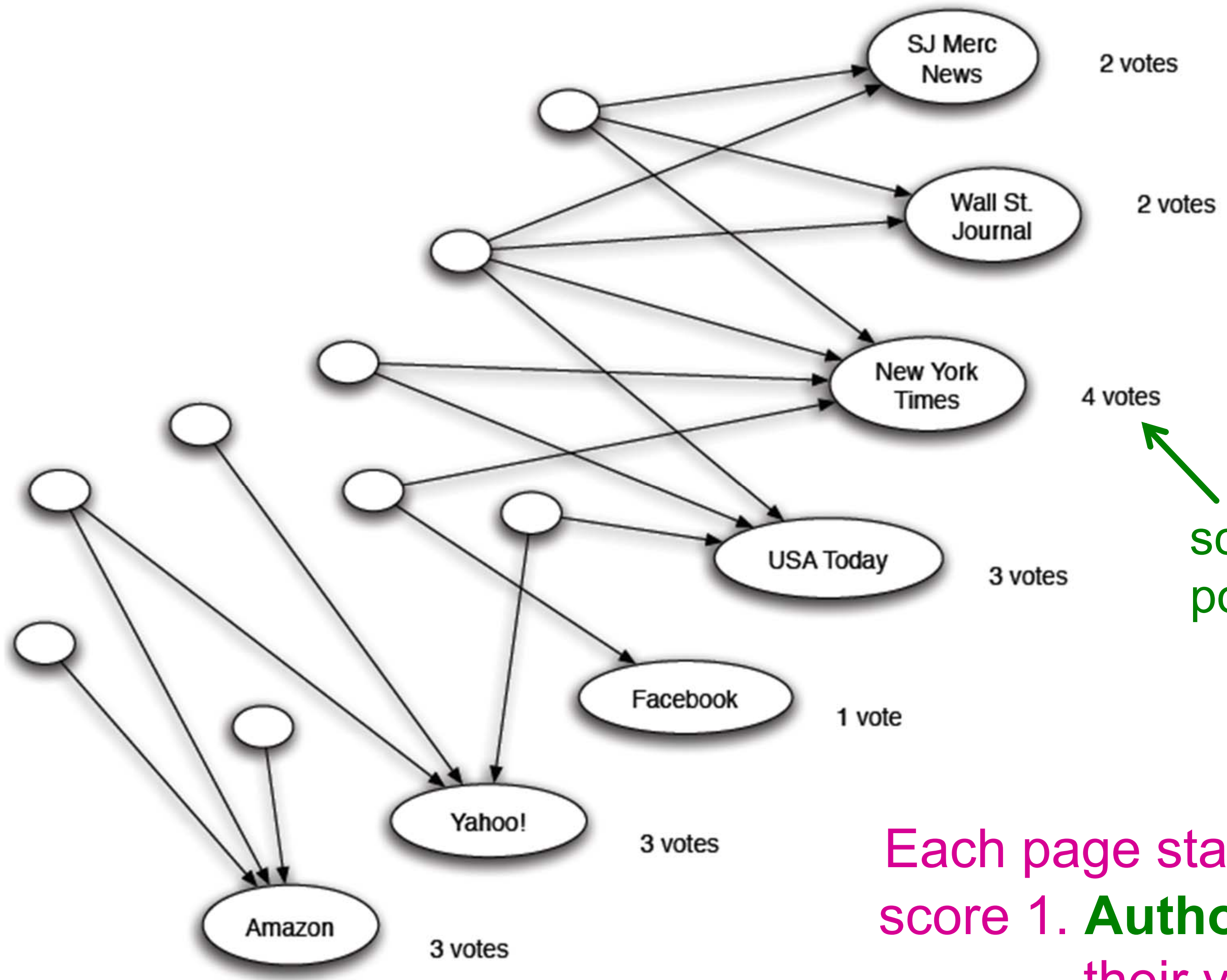
Counting in-links: Authority



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



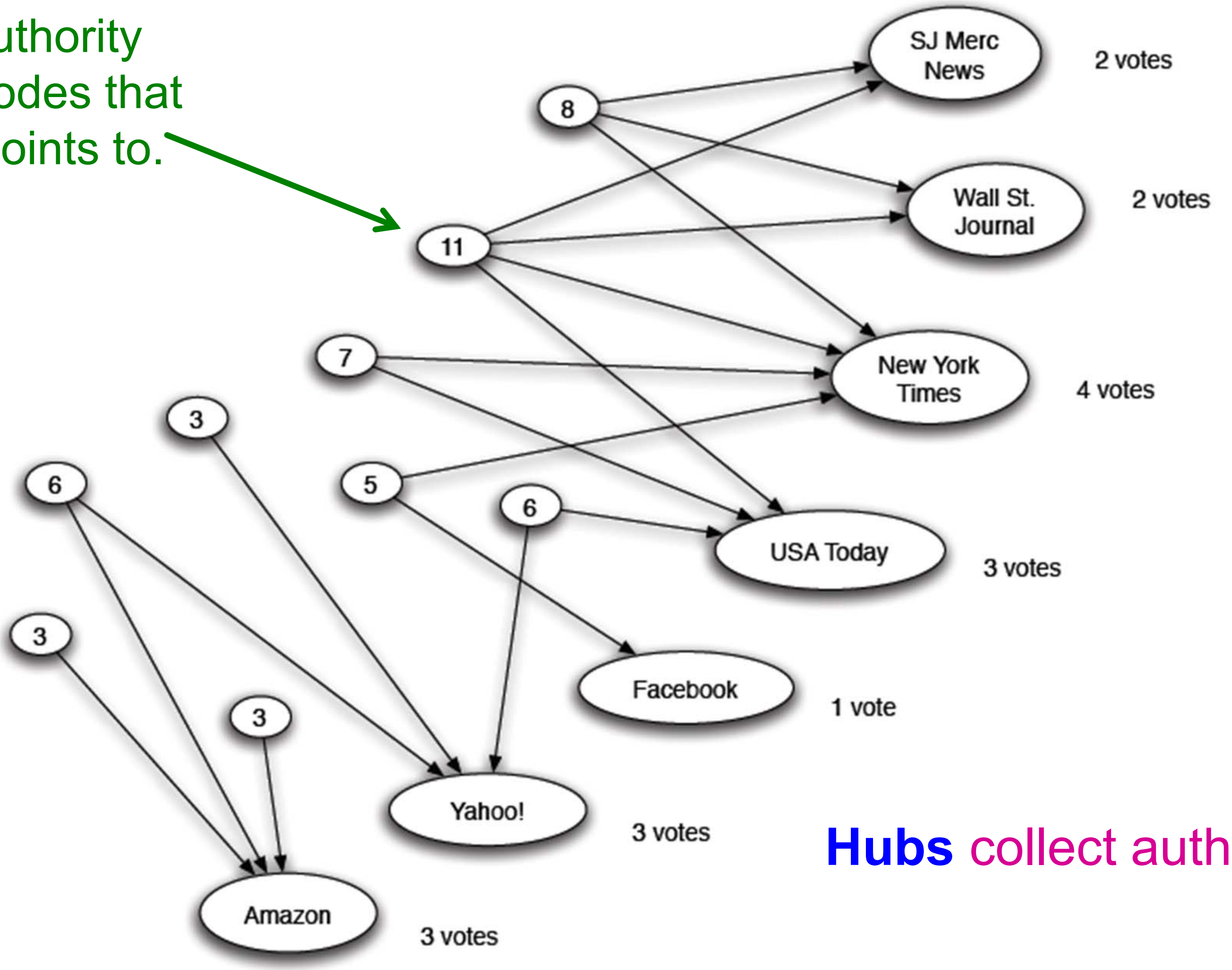
Sum of **hub** scores of nodes pointing to NYT.

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)

Expert Quality: Hub

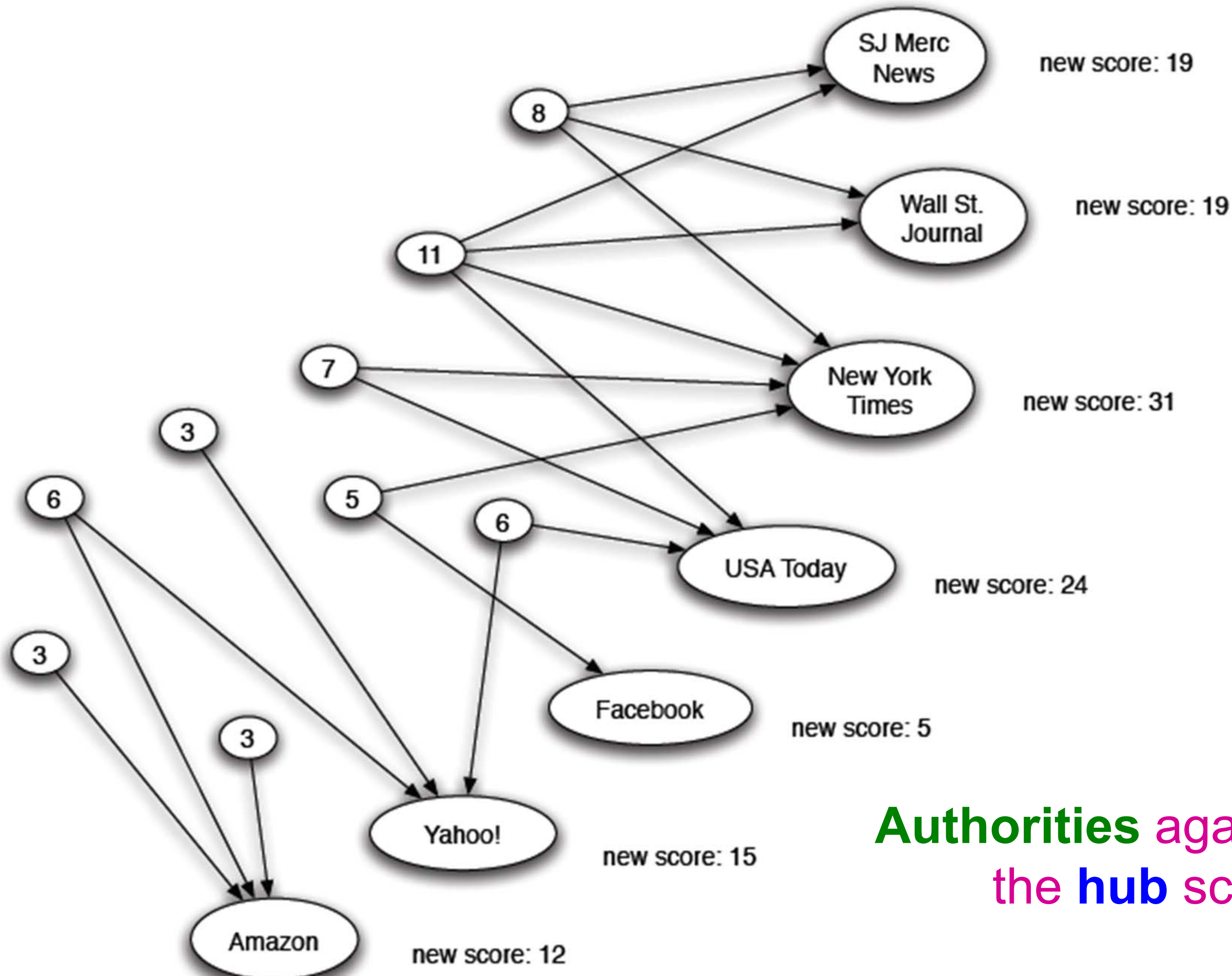
Sum of authority scores of nodes that the node points to.



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)

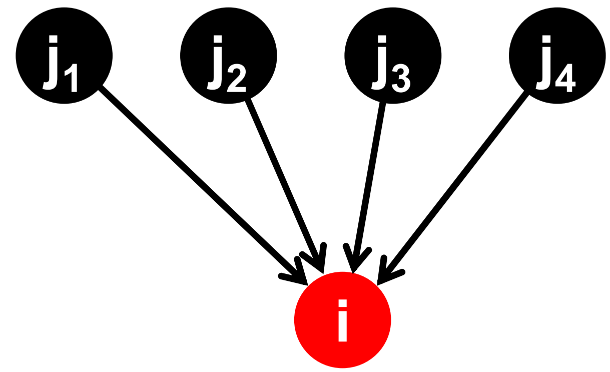
Mutually Recursive Definition

- **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 $\boldsymbol{h}$ and $\boldsymbol{a}$

Hubs and Authorities

[Kleinberg '98]

- **Each page *i* has 2 scores:**
  - Authority score: $\boldsymbol{a_i}$
  - Hub score: $\boldsymbol{h_i}$

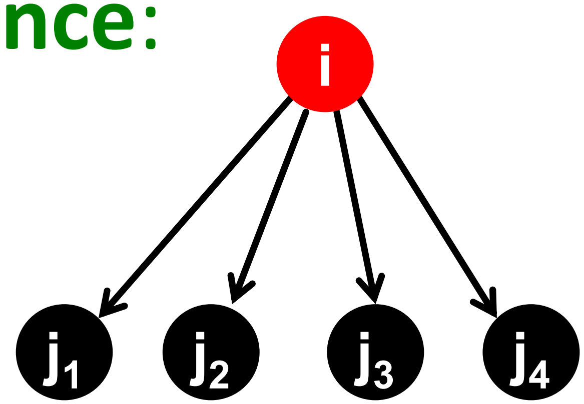


$$\boldsymbol{a_i = \sum_{j\to i} h_j}$$

HITS algorithm:

- Initialize: $a_j^{(0)} = 1/\sqrt{\mathrm{n}},\ \ \mathrm{h}_\mathrm{j}^{(0)} = 1/\sqrt{\mathrm{n}}$
- Then keep iterating until **convergence**:
  - $\forall \boldsymbol{i}$: Authority: $a_i^{(t+1)} = \sum_{j\to i} h_j^{(t)}$
  - $\forall \boldsymbol{i}$: Hub: $h_i^{(t+1)} = \sum_{i\to j} a_j^{(t)}$
  - $\forall \boldsymbol{i}$: Normalize:

$$\sum_i \left(a_i^{(t+1)}\right)^2 = 1, \sum_j \left(h_j^{(t+1)}\right)^2 = 1$$



$$\boldsymbol{h_i = \sum_{i\to j} a_j}$$

Hubs and Authorities

[Kleinberg '98]

- HITS converges to a single stable point
- **Notation:**
  - Vector $\boldsymbol{a} = (a_1 \dots, a_n)$, $\boldsymbol{h} = (h_1 \dots, h_n)$
  - Adjacency matrix $\boldsymbol{A}$ (n x n): $\boldsymbol{A_{ij}} = 1$ if $i{\rightarrow}j$
- Then $\boldsymbol{h_i} = \sum_{i \rightarrow j} \boldsymbol{a_j}$

  can be rewritten as $\boldsymbol{h_i} = \sum_j \boldsymbol{A_{ij}} \cdot \boldsymbol{a_j}$

  So: $\boldsymbol{h} = \boldsymbol{A} \cdot \boldsymbol{a}$
- Similarly, $\boldsymbol{a_i} = \sum_{j \rightarrow i} \boldsymbol{h_j}$

  can be rewritten as $\boldsymbol{a_i} = \sum_j \boldsymbol{A_{ji}} \cdot \boldsymbol{h_j} = \boldsymbol{A^T} \cdot \boldsymbol{h}$

Hubs and Authorities

- **HITS algorithm in vector notation:**
  - Set: $\boldsymbol{a_i} = \boldsymbol{h_i} = \frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}}$

  Repeat until convergence:
  - $\boldsymbol{h} = \boldsymbol{A} \cdot \boldsymbol{a}$
  - $\boldsymbol{a} = \boldsymbol{A^T} \cdot \boldsymbol{h}$
  - Normalize $\boldsymbol{a}$ and $\boldsymbol{h}$
- **Then:** $\boldsymbol{a} = \underbrace{\boldsymbol{A^T} \cdot (\underbrace{\boldsymbol{A} \cdot \boldsymbol{a}}_{\text{new } h})}_{\text{new } a}$

Convergence criterion:

$$\sum_i \left(h_i^{(t)} - h_i^{(t-1)}\right)^2 < \varepsilon$$

$$\sum_i \left(a_i^{(t)} - a_i^{(t-1)}\right)^2 < \varepsilon$$

a **is updated (in 2 steps):**

$$a = A^T(A\,a) = (A^T A)\,a$$

h **is updated (in 2 steps):**

$$h = A\,(A^T h) = (A\,A^T)\,h$$

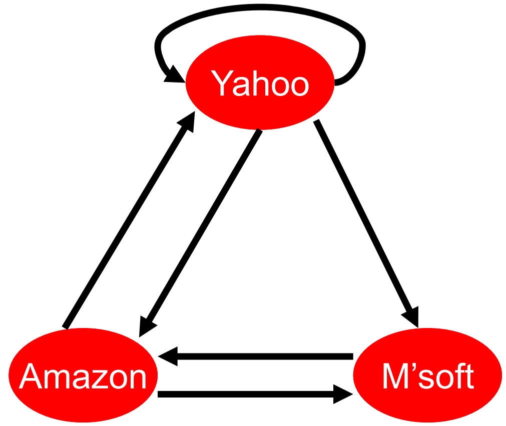
Repeated matrix powering

Existence and Uniqueness

- Under reasonable assumptions about **A**, HITS **converges to vectors *h** and *a****:
  - ***h**** is the **principal eigenvector** of matrix $\boldsymbol{A}\,\boldsymbol{A}^T$
  - ***a**** is the **principal eigenvector** of matrix $\boldsymbol{A}^T\boldsymbol{A}$

Example

$$A = \boxed{\begin{matrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{matrix}} \qquad A^T = \boxed{\begin{matrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{matrix}}$$



$$\begin{array}{lcllllll}
h(yahoo) & = & .58 & .80 & .80 & .79 & \cdots & .788 \\
h(amazon) & = & .58 & .53 & .53 & .57 & \cdots & .577 \\
h(m'soft) & = & .58 & .27 & .27 & .23 & \cdots & .211
\end{array}$$

$$\begin{array}{lcllllll}
a(yahoo) & = & .58 & .58 & .62 & .62 & \cdots & .628 \\
a(amazon) & = & .58 & .58 & .49 & .49 & \cdots & .459 \\
a(m'soft) & = & .58 & .58 & .62 & .62 & \cdots & .628
\end{array}$$

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**