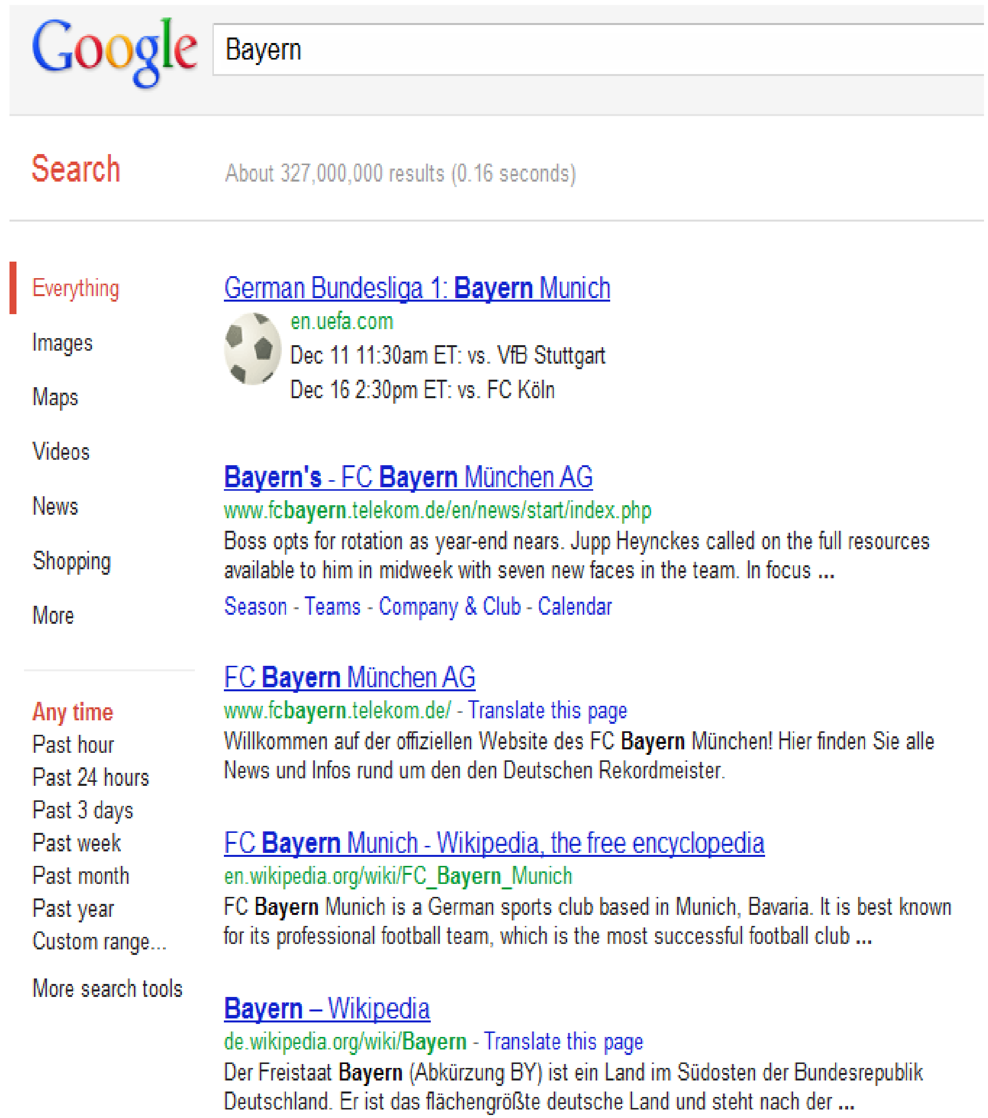
Link Analysis and Web Search (HITS)

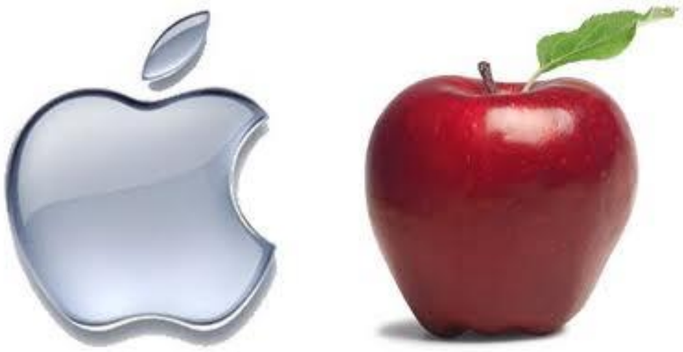
Searching the Web: The Problem of Ranking

- How does Google "know" what is the best answer?
- Search engines rank using automated methods that look at the Web itself
- Information intrinsic to the Web and its structure



A Hard Problem

- Information retrieval decades before the Web
  - newspaper articles, scientific papers, patents, legal abstracts
- Problems
  - Limited expressiveness of keywords
    - “Hildesheim”: town or university?
  - Synonymy: multiple ways to say the same thing
    - car, automobile, vehicle
  - Polysemy: multiple meanings for the same term

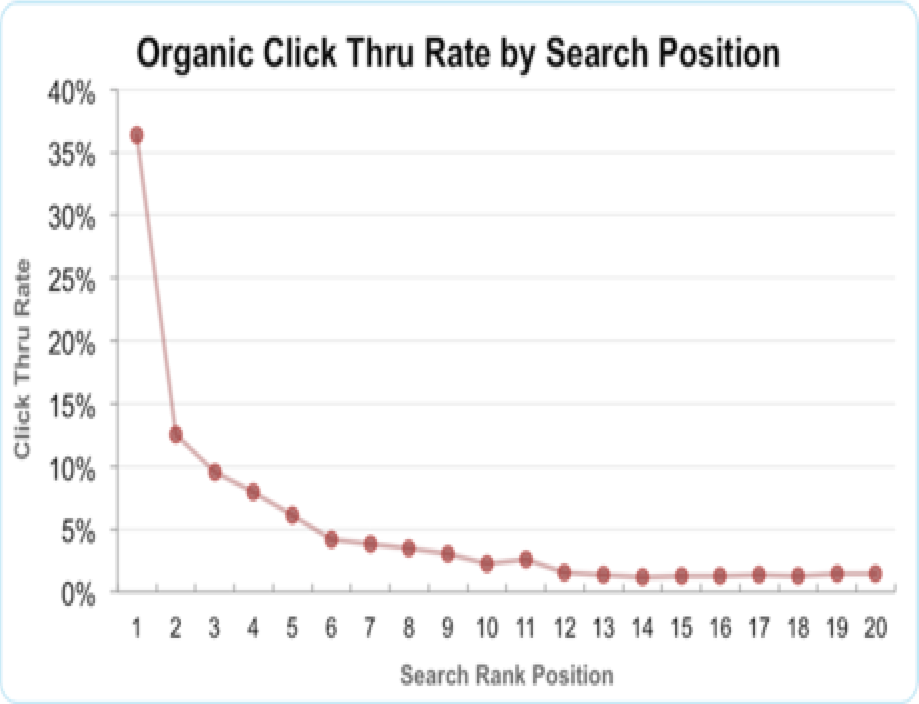


Dynamic Web Content

- Constantly-changing nature of Web content
- Example:
  - Search terms: “World Trade Center” on September 11, 2001
  - top results were pages about the building itself
- In response Google built specialized “News Search” which collect articles continuously
- Twitter fills in the spaces about real-time awareness

A Problem of Abundance

- Search engines find millions of documents relevant to a query
- Humans look only at few
- Which few should be shown?
- (Business models on this: next lecture)



Hyperlink-Induced Topic Search (HITS)

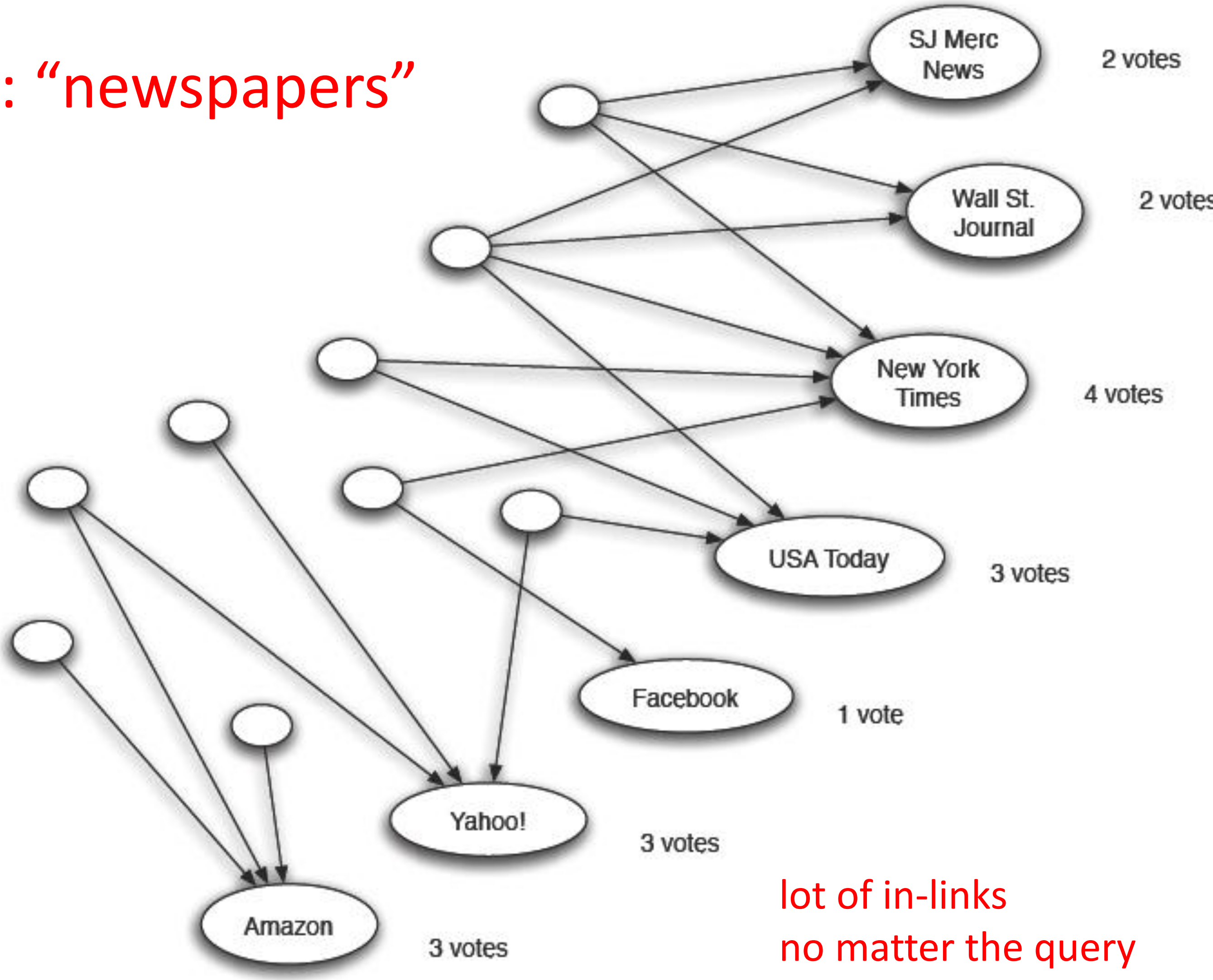
- HITS (aka hubs and authorities)
  - link analysis algorithm that ranks Web pages, by Jon Kleinberg
  - Precursor to PageRank
- Hubs serve as compilations (catalogs, lists) of information leading to authoritative pages
- Let’s see how it works…

Voting by In-Links

- Links are essential to ranking
  - Assume that page P is the best result of query Q
  - When a page X is relevant to a query Q, P is among the pages X links to
- Each link may have many possible meanings
  - may convey criticism
  - may be a paid advertisement
  - in aggregate many links represent collective endorsement
- Method:
  - first collect a large sample of pages relevant to the query (text-based IR)
  - pages in this sample “vote” through their links

Example of Voting by In-Links

Query: "newspapers"

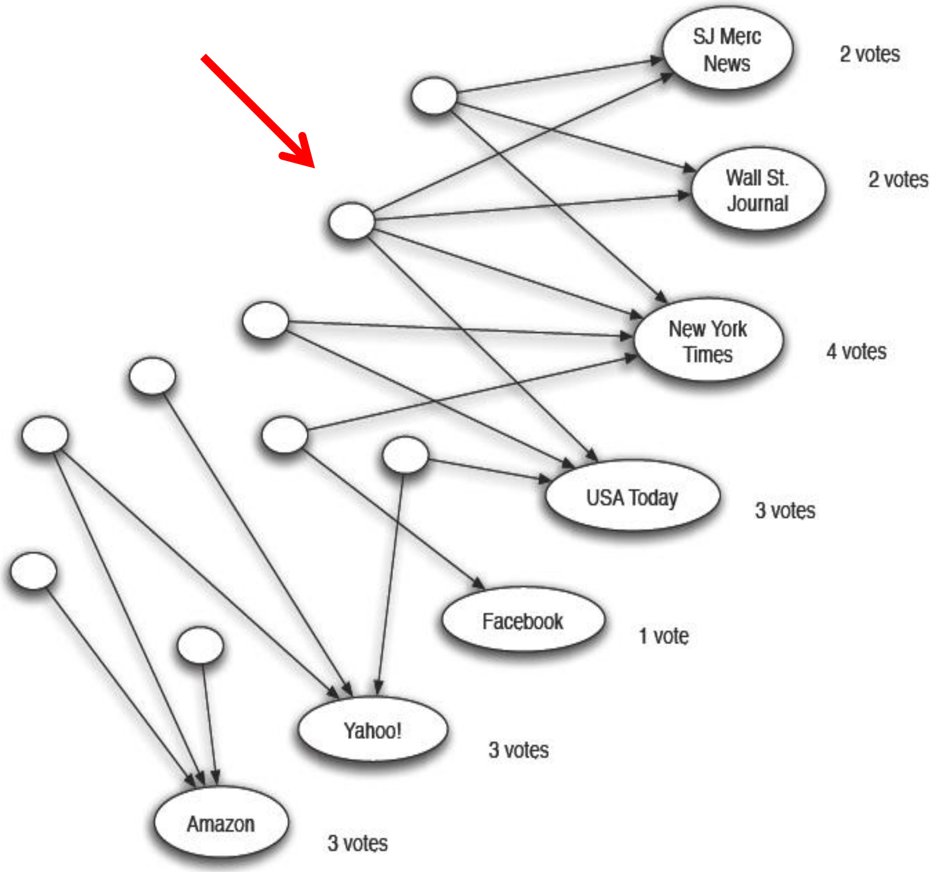


Prominent newspapers

lot of in-links
no matter the query

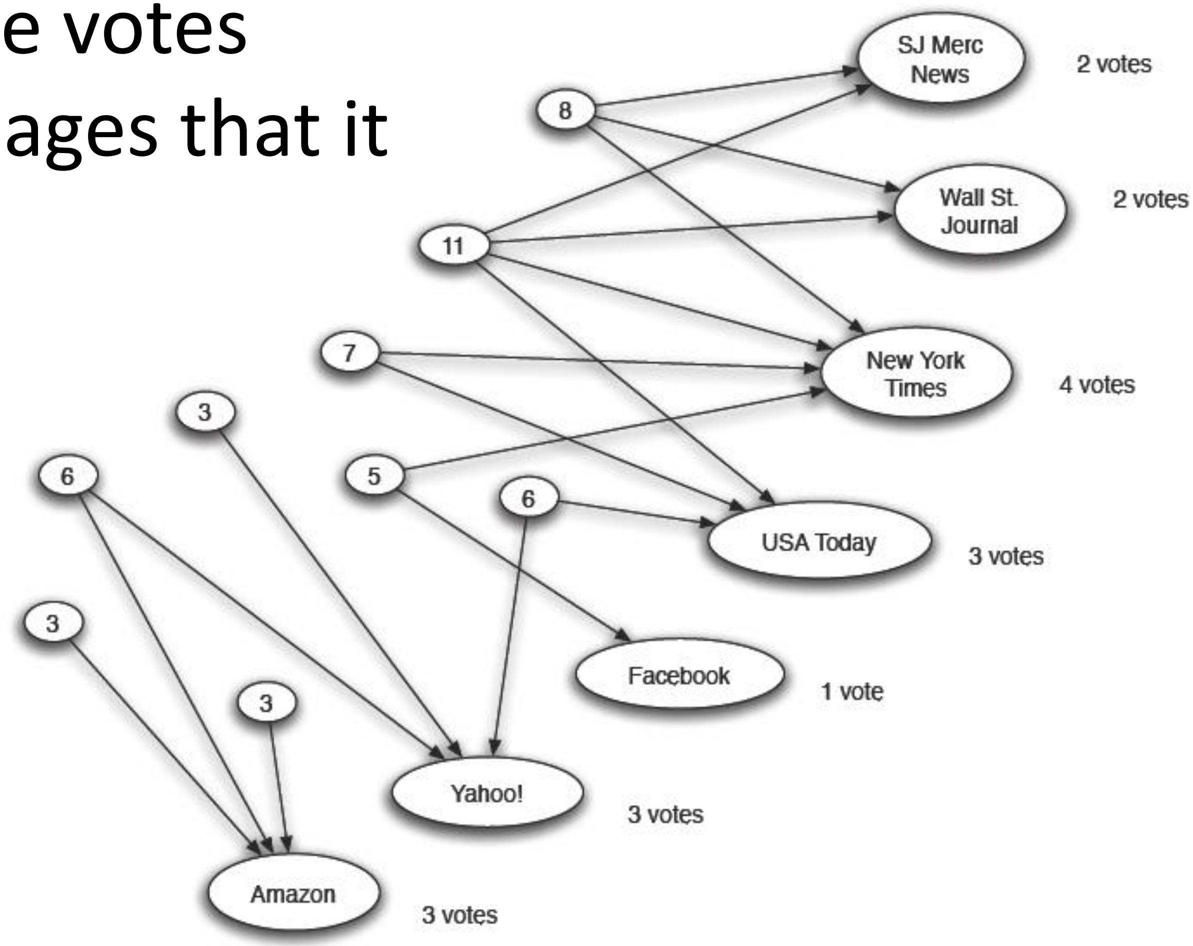
A List-Finding Technique

- Among the pages casting votes, a few vote for many of the authoritative pages (those receiving a lot of votes)
- Pages that compile lists of resources relevant to the query topic



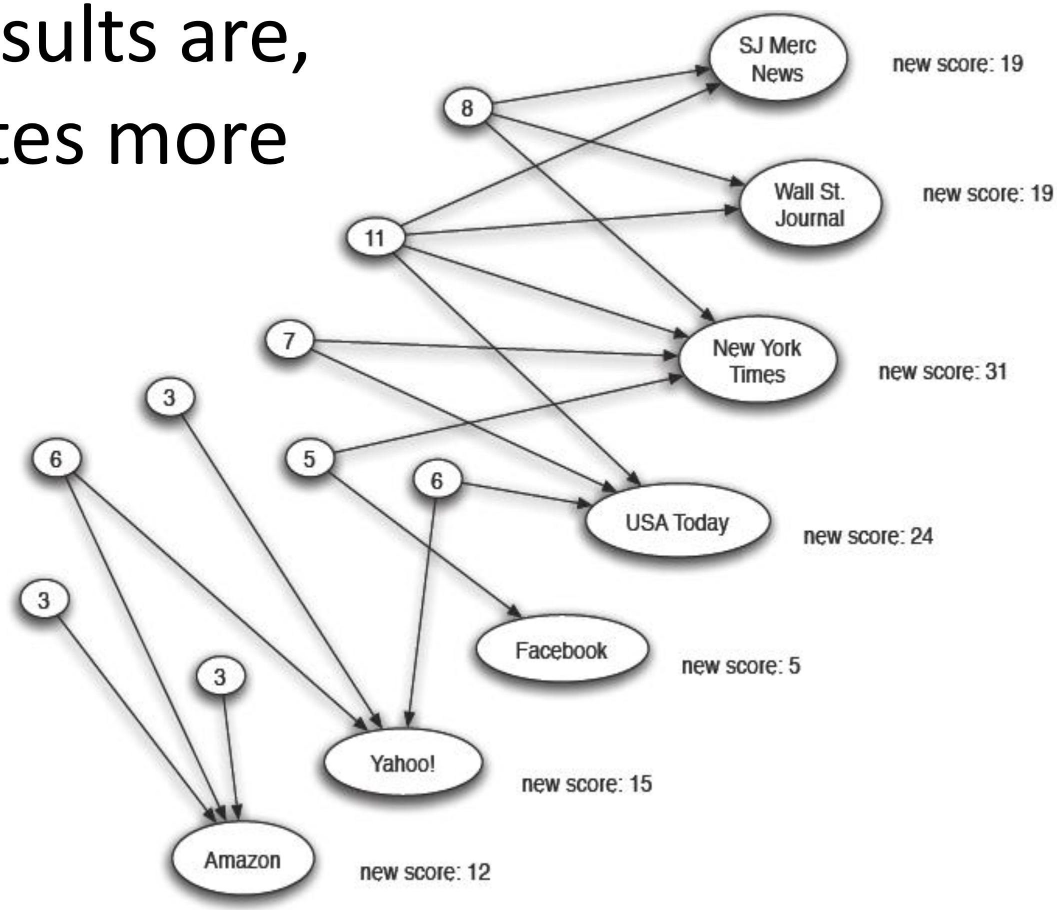
A List-Finding Technique

- A page's value as a list is equal to the sum of the votes received by all pages that it voted for



The Principle of Repeated Improvement

- If lists link to good results are, then weight their votes more heavily
- Cast the votes again
  - Each page's vote a weight equal to its value as a list



The Principle of Repeated Improvement

- Why stop here?
- If we have better votes on the authorities, we can use them to get better scores for the lists
- The process can go back and forth forever

Hubs and Authorities

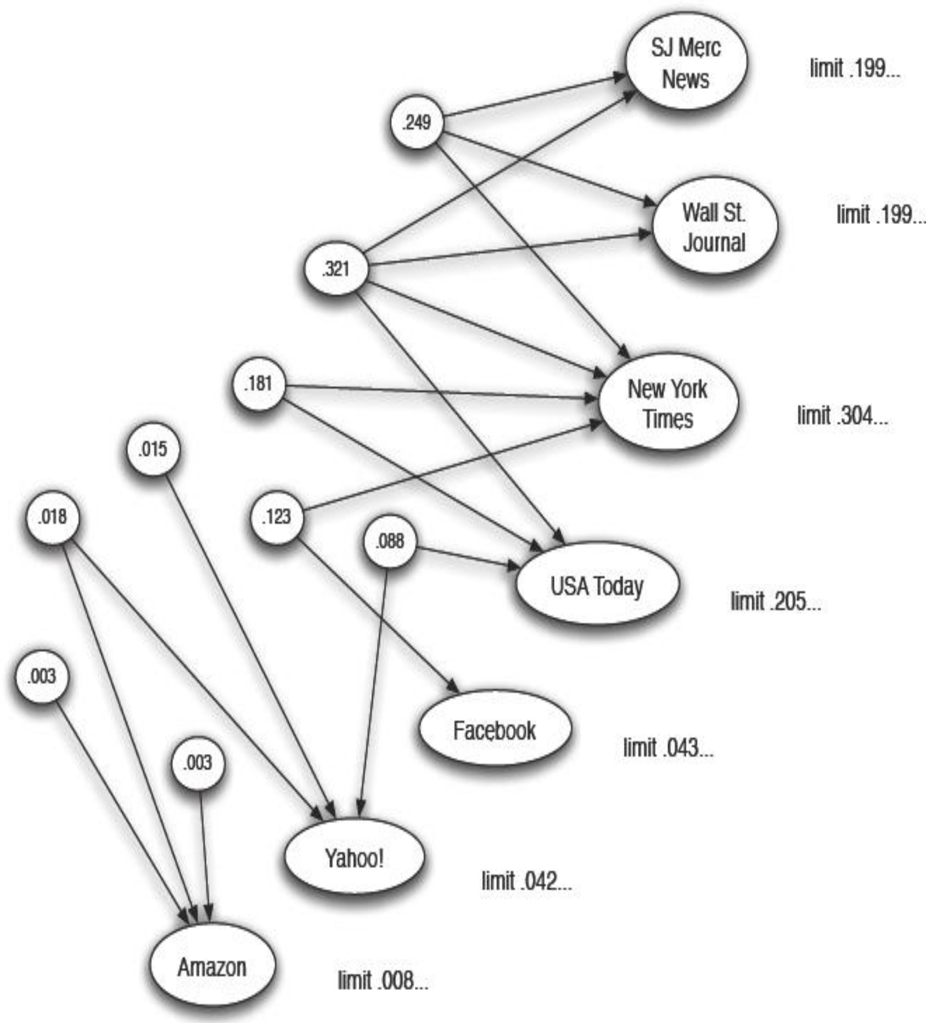
- For each page p
  - auth(p): its value as a potential authority
  - hub(p): its value as a potential hub
- Authority Update Rule:
  - For each page p, update auth(p) to be the sum of the hub scores of all pages that point to it
- Hub Update Rule:
  - For each page p, update hub(p) to be the sum of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

HITS Algorithm

- Start with all hub scores and all authority scores equal to 1
- Choose a number of steps k
- 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: divide down each authority score by the sum of all authority scores, and divide down each hub score by the sum of all hub scores

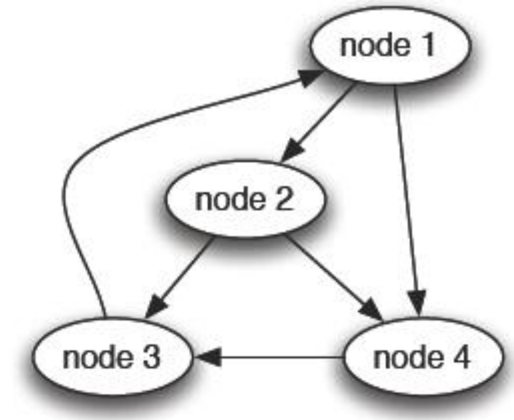
Convergence of HITS

- What happens for larger and larger values of k?
- Normalized values actually converge to limits as k goes to infinity
  - results stabilize so that continued improvement leads to smaller and smaller changes
  - we reach the same limiting values no matter what we choose as the initial hub and authority values



Spectral Analysis of HITS

- Adjacency Matrices
  - n × n matrix M
  - M_{ij} = 1 if link i-> j, M_{ij} = 0 otherwise
- Hub/Authority Scores
  - hub (h) and authority (a) vectors n x 1
- How to write the Authority Update and Hub Update Rules as matrix-vector multiplications



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

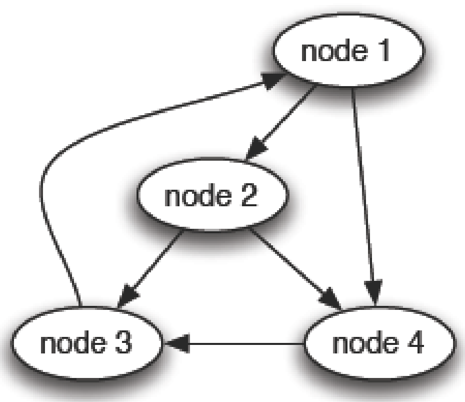
Hub and Authority Update Rules as Matrix-Vector Multiplication

- For a node i, its hub score h_i is updated to be the sum of a_j over all nodes j to which i has an edge

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

- Matrix-vector multiplication form

$$h \leftarrow Ma$$



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 2 \\ 4 \end{bmatrix}$$

Hub and Authority Update Rules as Matrix-Vector Multiplication

- For a node i, its authority score a_i is updated to be the sum of h_j over all nodes j that have an edge to i

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

- Matrix-vector multiplication form using a matrix rows and columns are interchanged (transpose of the matrix)

$$a \leftarrow M^T h$$

k-step Hub-Authority Computation

- Initial vectors of authority and hub scores

$a^{\langle 0 \rangle}$ and $h^{\langle 0 \rangle}$

- Authority and hub scores after k applications of the Authority and then Hub Update Rules

$a^{\langle k \rangle}$ and $h^{\langle k \rangle}$

- Apply matrix-multiplication formulas:

$$a^{\langle 1 \rangle} = M^T h^{\langle 0 \rangle}$$

$$h^{\langle 1 \rangle} = M a^{\langle 1 \rangle} = M M^T h^{\langle 0 \rangle}$$

k-step Hub-Authority Computation

- In the second step (k=2)

$$a^{\langle 2 \rangle} = M^T h^{\langle 1 \rangle} = M^T M M^T h^{\langle 0 \rangle}$$

$$h^{\langle 2 \rangle} = M a^{\langle 2 \rangle} = M M^T M M^T h^{\langle 0 \rangle} = (M M^T)^2 h^{\langle 0 \rangle}$$

- In the third step (k=3)

$$a^{\langle 3 \rangle} = M^T h^{\langle 2 \rangle} = M^T M M^T M M^T h^{\langle 0 \rangle} = (M^T M)^2 M^T h^{\langle 0 \rangle}$$

$$h^{\langle 3 \rangle} = M a^{\langle 3 \rangle} = M M^T M M^T M M^T h^{\langle 0 \rangle} = (M M^T)^3 h^{\langle 0 \rangle}$$

What is the recursive rule?

k-step Hub-Authority Computation

- In the k-th step

$$a^{\langle k \rangle} = (M^T M)^{k-1} M^T h^{\langle 0 \rangle}$$

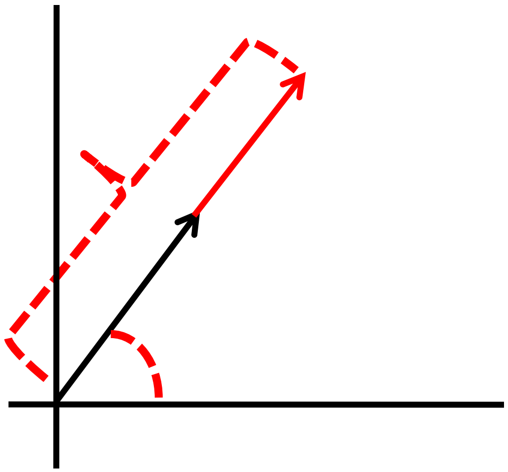
$$h^{\langle k \rangle} = (M M^T)^k h^{\langle 0 \rangle}$$

- Authority and hub vectors are the results of multiplying an initial vector by larger and larger powers of M^TM and MM^T respectively

Does this process converge to stable values?

Multiplication in terms of eigenvectors

- Magnitude of the hub and authority values tend to grow with each update
- They will only converge when we take normalization into account
- The directions of the hub and authority vectors that are converging



Multiplication in terms of eigenvectors

- There are (normalizing) constants c and d s.t.:

$$\frac{h^{\langle k\rangle}}{c^k} \text{ and } \frac{a^{\langle k\rangle}}{d^k}$$

  converge to limits as *k goes to infinity*
- Taking the recursive formula:

$$\frac{h^{\langle k\rangle}}{c^k} = \frac{(MM^T)^k h^{\langle 0\rangle}}{c^k}$$

- $h^{<k>}$ converges to limit $h^{<*>}$ if the direction does not change when multiplied with MM^T (but magnitude may change by a factor c)

$$(MM^T)h^{\langle *\rangle} = ch^{\langle *\rangle}$$

Multiplication in terms of eigenvectors

- Q: When a vector **v** doesn't change direction when multiplied by a given matrix **X**?
- A: When **v** is an eigenvector of **X**
  - $\mathbf{X}\ \mathbf{v} = \lambda\mathbf{X}$
  - A solution of: $\det(\mathbf{X} - \lambda\mathbf{I}) = 0$

- **Definition**: The eigenvectors of a square matrix are the non-zero vectors that, after being multiplied by the matrix, remain parallel to the original vector

- It follows that $h^{<*>}$ has to be an eigenvector of MM^T

Convergence of the hub-authority

- We have to prove that the direction of $h^{<k>}$ (normalized: $h^{<k>}/c^k$) converges to an eigenvector of MM^T
  - MM^T is symmetric => has n eigenvectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (assume w.l.g. that: $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$)
- $h^{<k>} = (MM^T)^k h^{<0>}$
- $(MM^T)^k h^{<0>} = (MM^T)^k (q_1 \mathbf{v_1} + \ldots + q_n \mathbf{v_n}) =$
  $q_1 (MM^T)^k \mathbf{v_1} + \ldots + q_n (MM^T)^k \mathbf{v_n} =$
  $q_1 (\lambda_1)^k \mathbf{v_1} + \ldots + q_n (\lambda_n)^k \mathbf{v_n}$

Convergence of the hub-authority

- $h^{<k>} = (\lambda_1)^k q_1 \mathbf{v_1} + \ldots + (\lambda_n)^k q_n \mathbf{v_n}$
- Assume $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$
- $h^{<k>} / (\lambda_1)^k = q_1 \mathbf{v_1} + q_1 (\lambda_2/\lambda_1)^k \mathbf{v_2} + \ldots + (\lambda_n/\lambda_1)^k q_n \mathbf{v_n}$
- As k goes to infinity, every term except the first goes to 0
- Therefore, $h^{<k>} / (\lambda_1)^k$ converges to $q_1 \mathbf{v_1}$
- Remaining to prove:
  - Relax assumption $|\lambda_1| > |\lambda_2|$
  - Proof regardless of initial vector $h^{<0>}$
  - See book: pages 423-424