Link Analysis: HITS & Page Rank

Advanced Social Computing

Department of Computer Science
University of Massachusetts, Lowell
Spring 2020

Hadi Amiri
hadi@cs.uml.edu

UMASS LOWELL

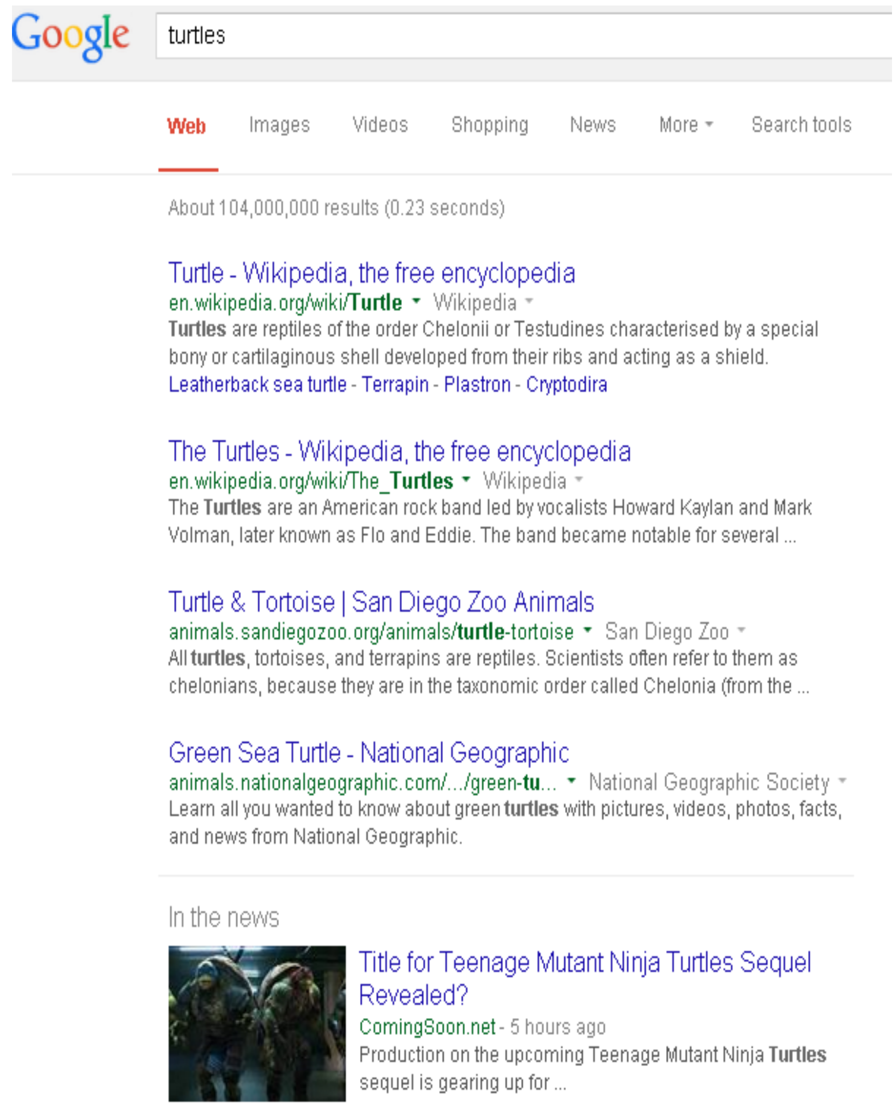
Ranking Problem

- We already know that there is considerable variation in the connectivity and structure of nodes in networks.

- How can we find nodes that are potentially more "important" or "authoritative" than others?

Ranking Problem- Cnt.

- Web Ranking Problem:
  - Given the Web and a query, rank Web pages with respect to the query such that the most relevant pages to the query appear higher in the list.



Lecture Topics

- **HITS**
- Spectral Analysis of HITS
- Page Rank
- Spectral Analysis of Page Rank

HITS

- Hyperlink-Induced Topic Search (HITS)
  - A Link analysis algorithm for ranking nodes.
- Links are essential for ranking
  - **In-links** could be considered as endorsements!
- In aggregate, if a node receives many links from other (important) nodes, then its is receiving **collective endorsement**!

Source: Authoritative sources in a hyperlinked environment. Kleinberg, Jon. Journal of the ACM (JACM). 1999.

HITS- Cnt.

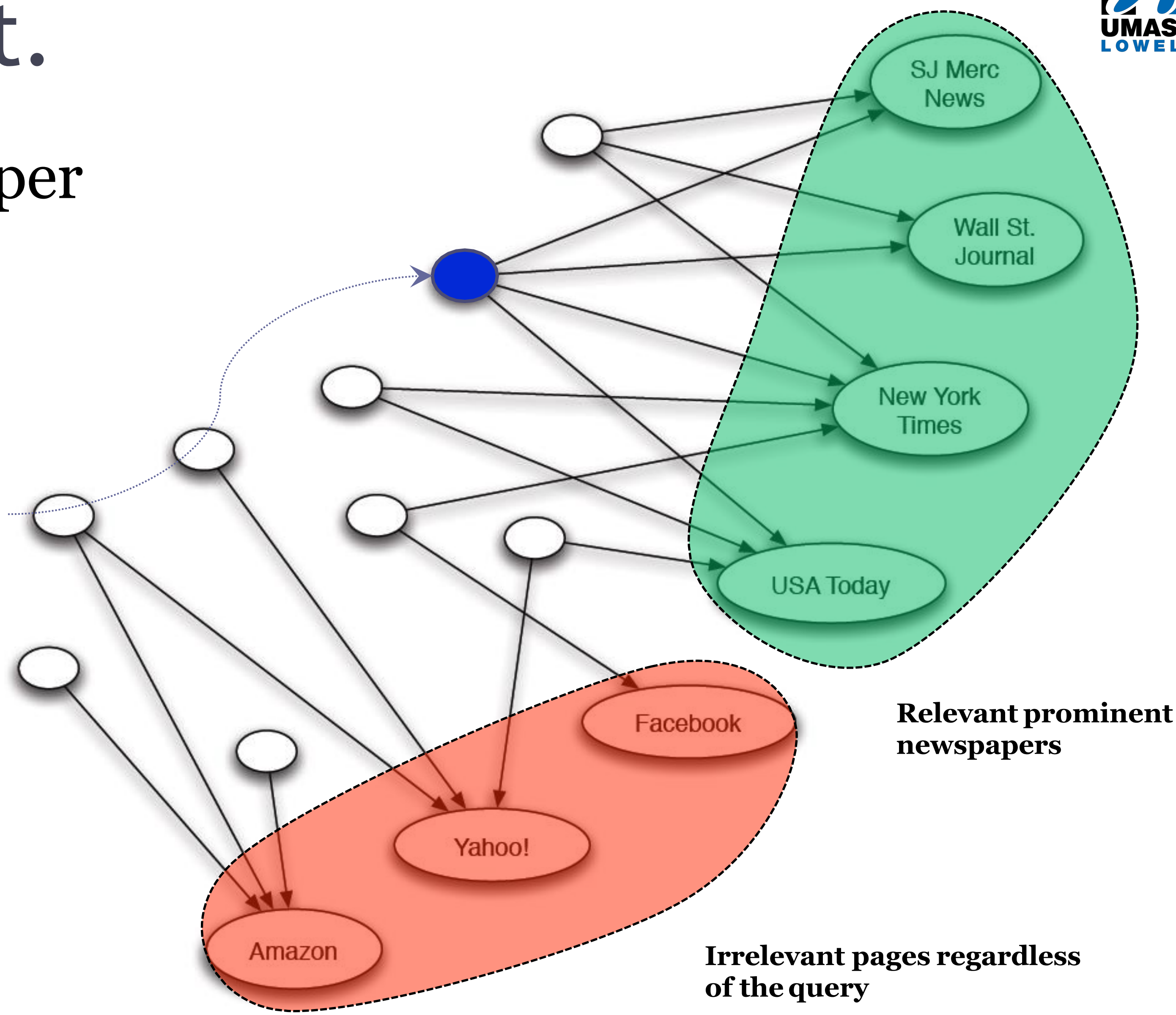
- How to operationalize such endorsement process?
  - Collect a large sample of pages relevant to the given query (e.g. “newspapers”)
    - Use text-based Information Retrieval
  - Pages in this sample “vote” / “endorse” through their links
    - A page is more important if it receives more votes (endorsement or in-links)

HITS- Cnt.

- Q: newspaper

Experts vote for many authoritative pages!

- these pages may have some sense of where the good answers are
- Score them highly



HITS- Cnt.

- Interesting Web pages fall into two categories:

1. **Authorities** that are pages containing relevant information
   - Newspapers homepages
   - Universities homepages

2. **Hubs** are pages that link to authorities
   - Lists of newspapers
   - Directories

HITS- Cnt.

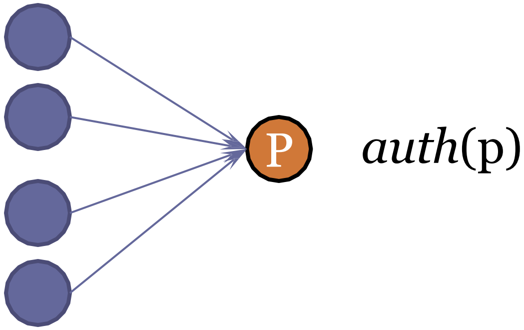
- A **good hub**?
  - links to many good authorities

- A **good authority**?
  - is linked from many good hubs

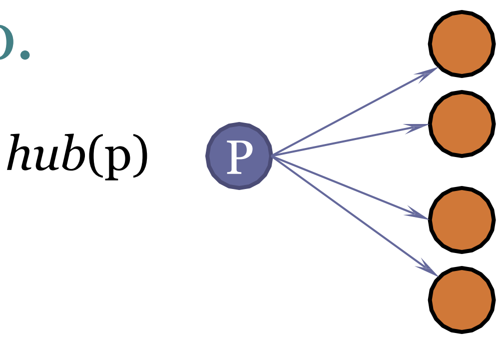
- We use two scores for each node
  - **Hub score** and **Authority score**

HITS- Cnt.

- **Authority Score**:
  - For each page p is the sum of the hub scores of all pages that point to it.



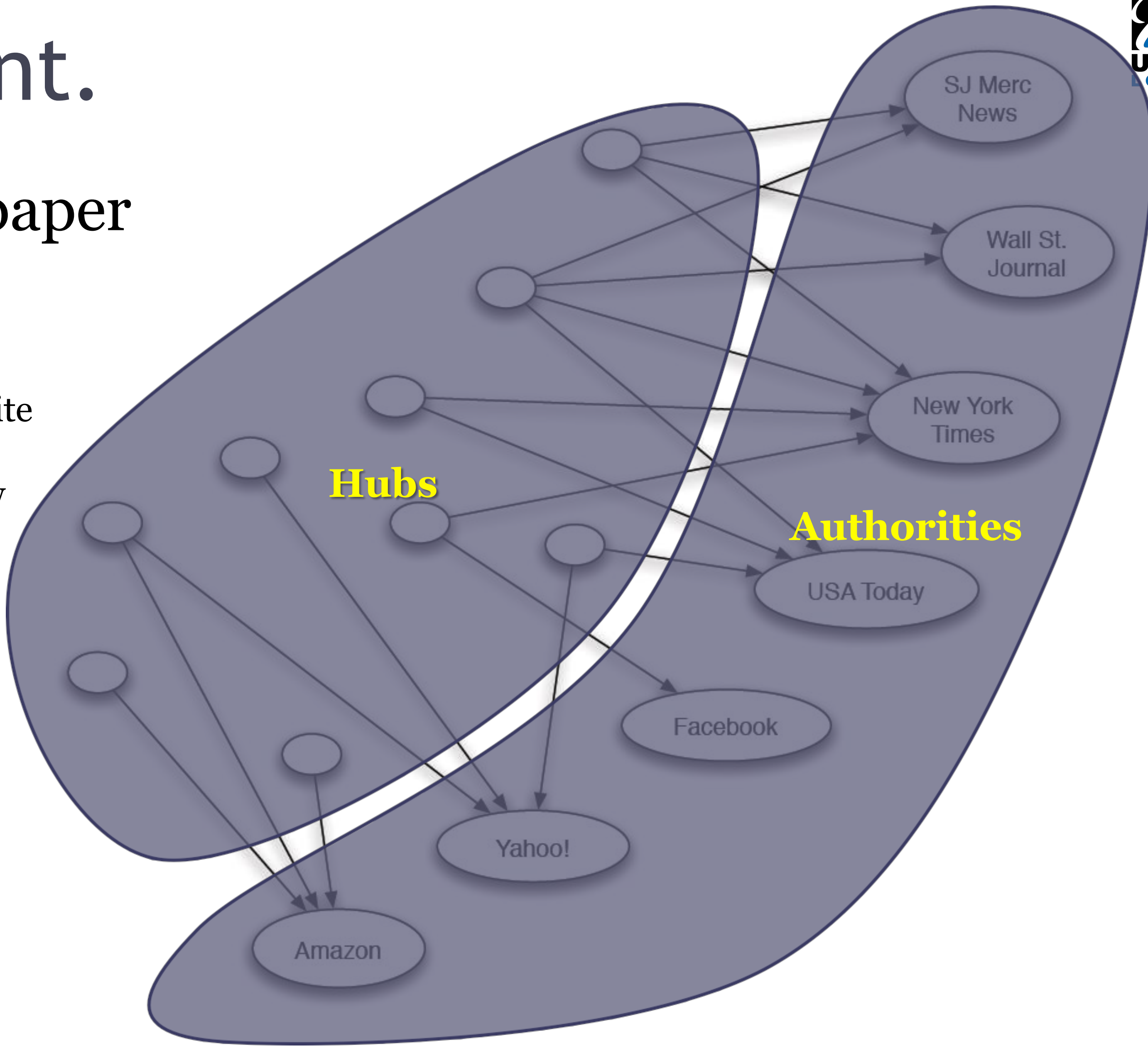
- **Hub Score**:
  - For each page p is the sum of the authority scores of all pages that it points to.



HITS- Cnt.

- Q: newspaper

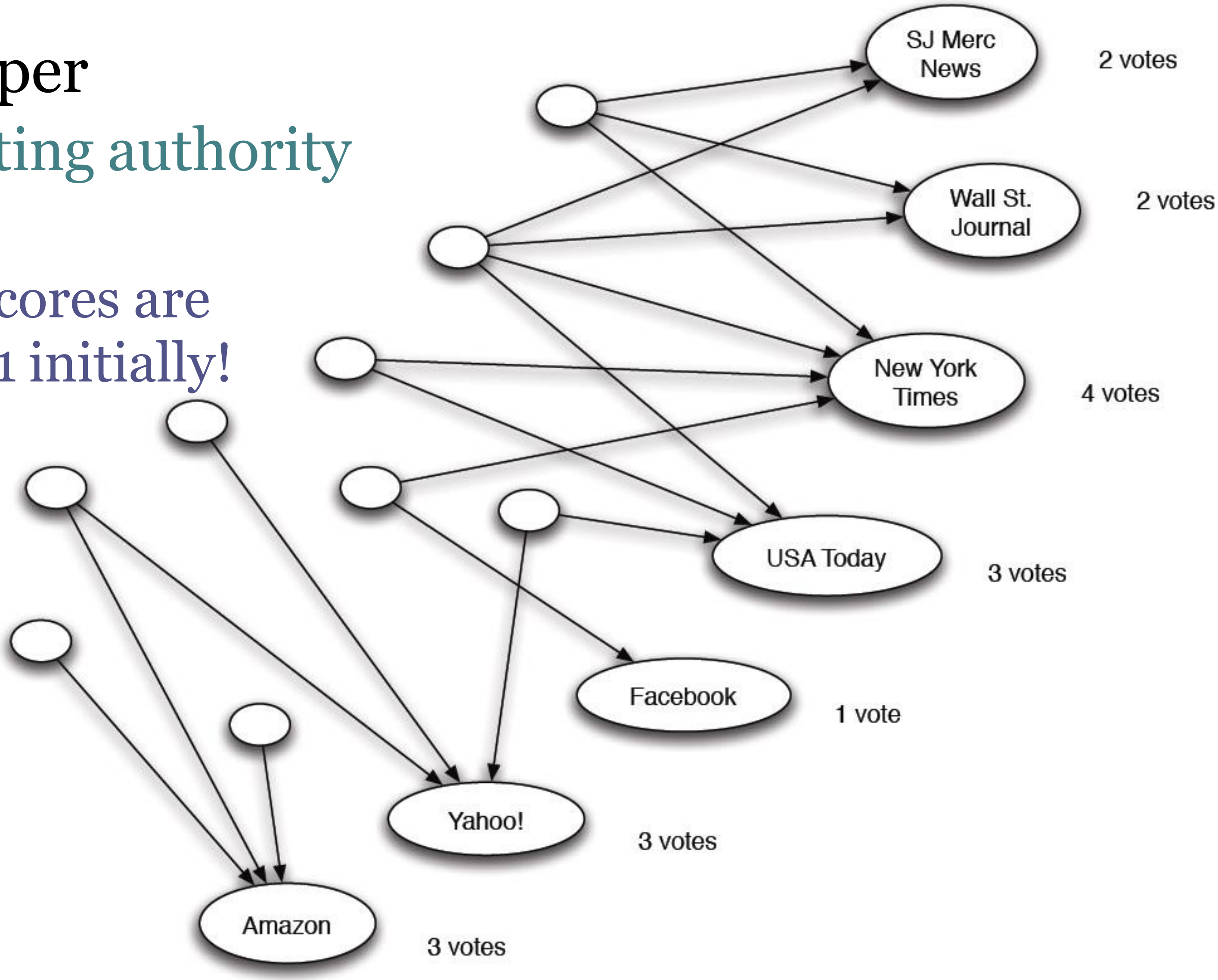
Let's assume a bipartite graph with hub and authority sets for now



Note that in practice the graph is not bipartite

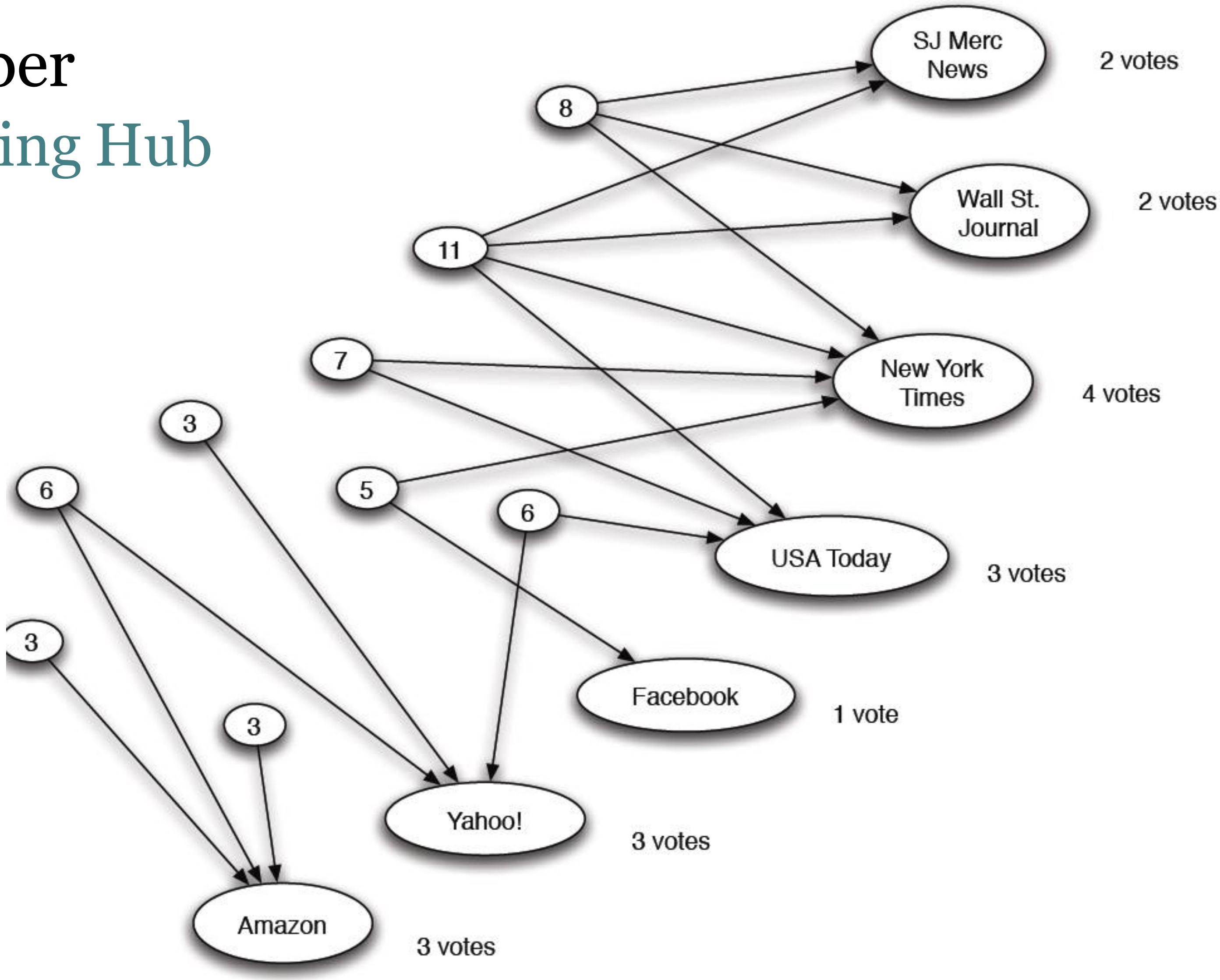
HITS- Cnt.

- Q: newspaper
  1. Computing authority scores
     - Hub scores are set to 1 initially!



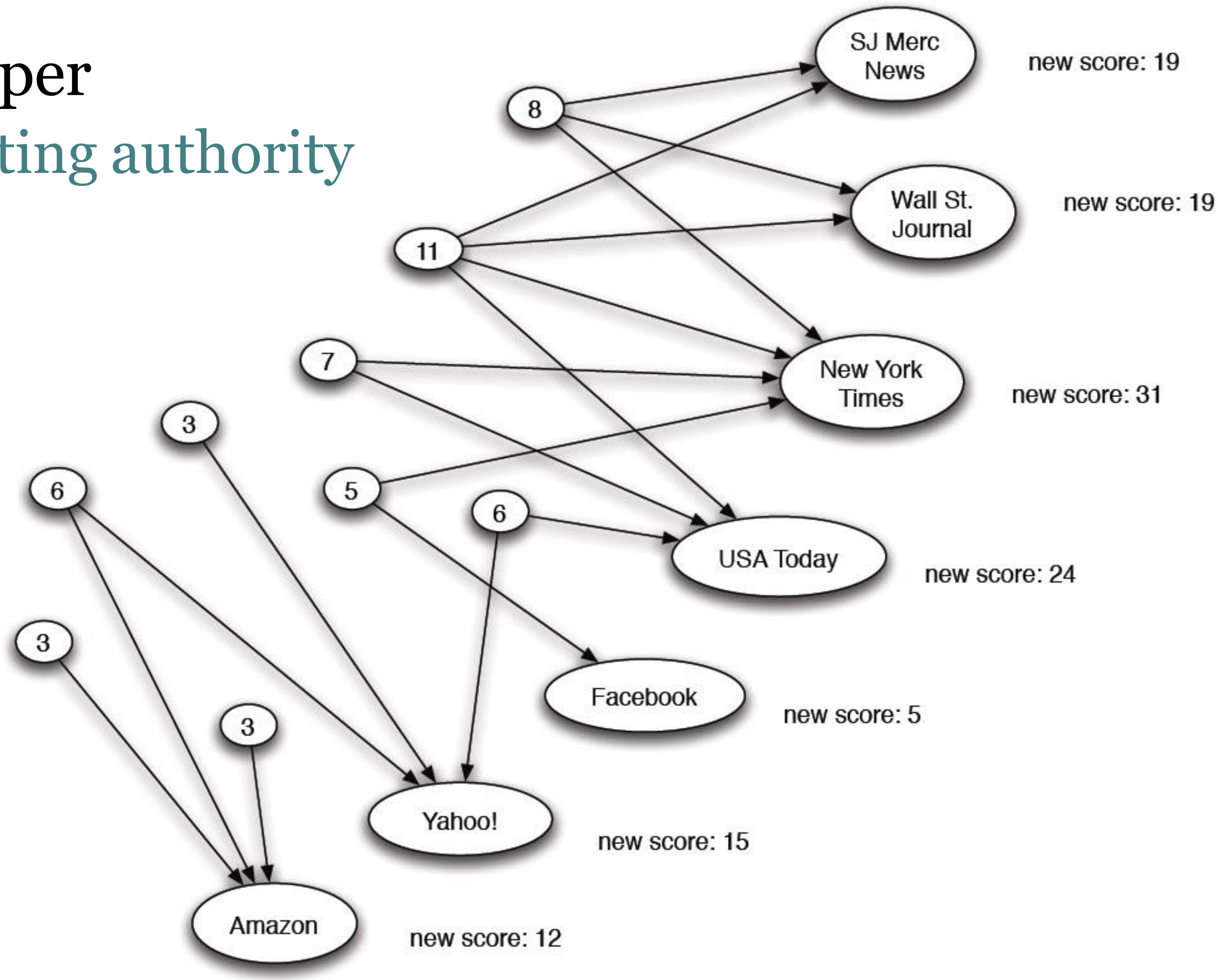
HITS- Cnt.

- Q: newspaper
  2. Computing Hub scores



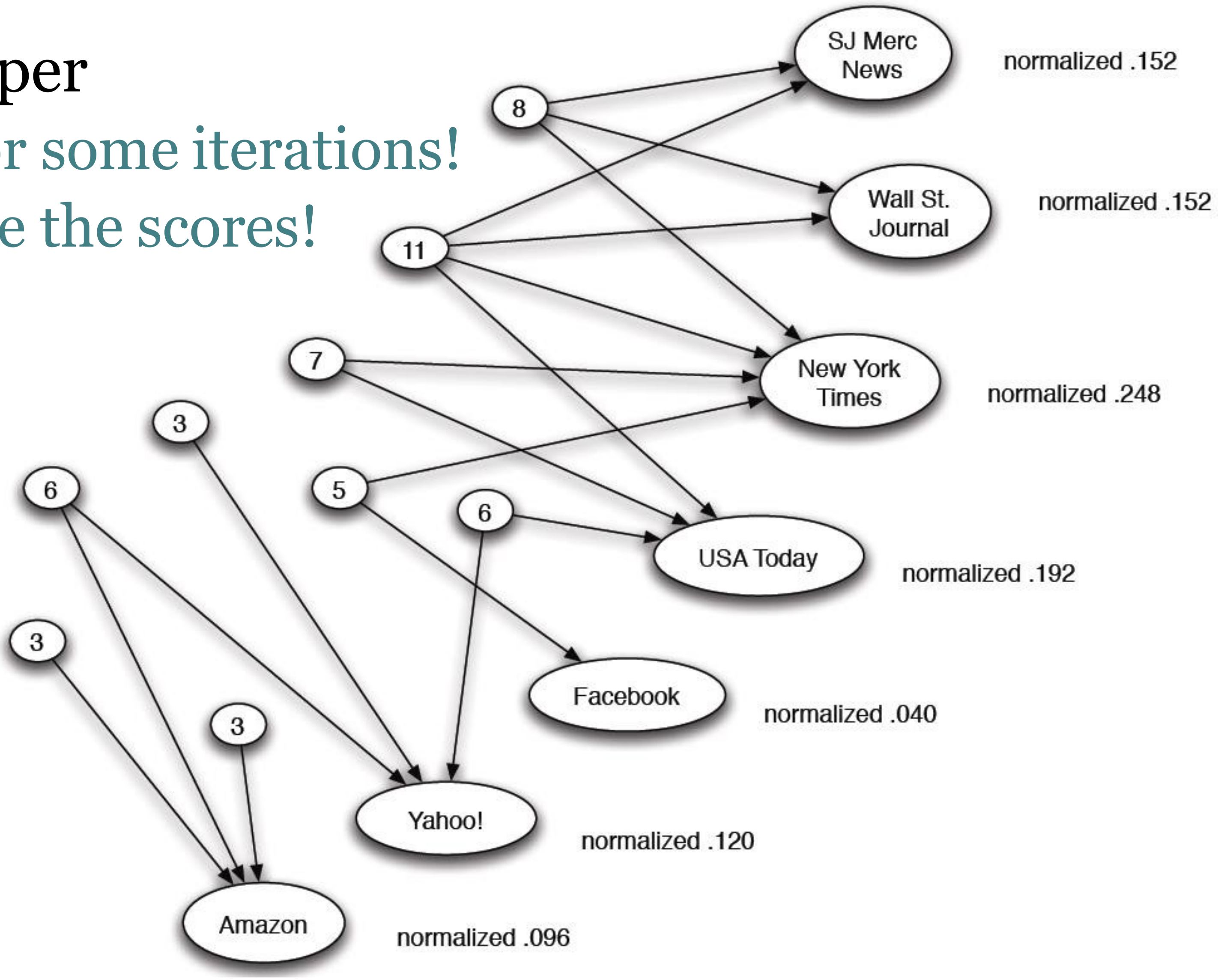
HITS- Cnt.

- Q: newspaper
  3. Computing authority scores



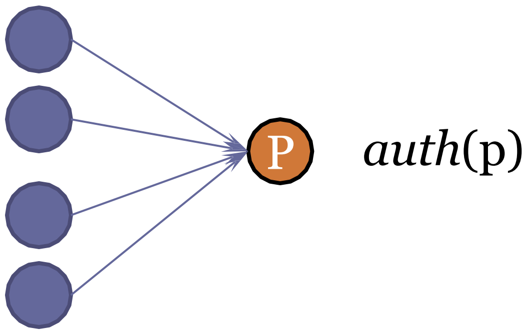
HITS- Cnt.

- Q: newspaper
    - Repeat for some iterations!
    - Normalize the scores!

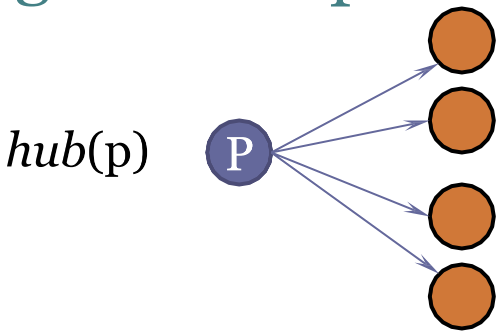


HITS- Cnt.

- **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- Cnt.

Algorithm

1. Set all hub scores and authority scores to 1.
2. Choose a number of steps k.
3. Perform a sequence of k hub-authority updates:
   1. First apply the Authority Update Rule to the current set of scores.
   2. Then apply the Hub Update Rule to the resulting set of scores.
4. Normalize authority and hub scores

HITS- Cnt.

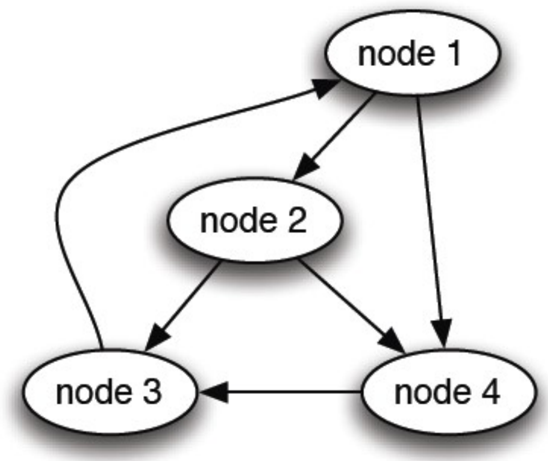
- What happens if we run HITS for larger and larger values of *k*?
  - The normalized values **converge** to limits as *k* goes to infinity!
    - Values stabilize; further updates lead to smaller and smaller changes in the values we observe!

Lecture Topics

- HITS
- **Spectral Analysis of HITS**
- Page Rank
- Spectral Analysis of Page Rank

Spectral Analysis of HITS

- Let $M_{ij} \in n \times n$
  - denote the adjacency matrix of our Web page sample!
    - If there is a directed edge from page i to page j
      - $M_{ij} = 1$
    - Otherwise
      - $M_{ij} = 0$



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

Spectral Analysis of HITS- Cnt.

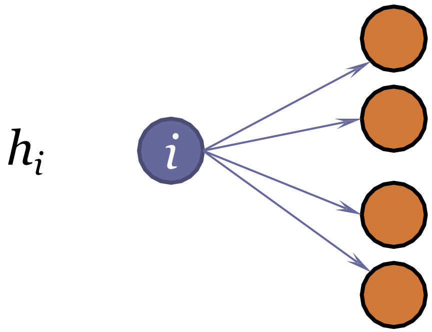
- Let $M_{ij} \in n \times n$
  - denote the adjacency matrix of our Web page sample!
- Represent Hub and Authority scores by two n-dimension vectors
  - Hub: $h \in n \times 1$
    - h_i represents the hub score of node i
  - Authority: $a \in n \times 1$
    - a_i represents authority score of node i

Spectral Analysis of HITS- Cnt.

- **Hub Update Rule**:
  - For each page i, update h_i to be the sum of the authority scores of all pages that it points to.

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

$$h \leftarrow Ma.$$

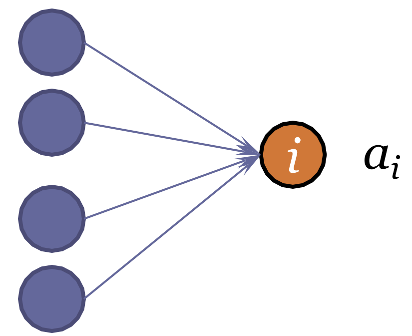


Spectral Analysis of HITS- Cnt.

- **Authority Update Rule**:
  - For each page i, update a_i to be the sum of the hub scores of all pages that point to it.

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

$$a \leftarrow M^T h.$$



Spectral Analysis of HITS- Cnt.

- Unwinding the *k*-step hub-authority computation
  - Let
    - $a^{<o>}$ initial vector of authority scores
    - $h^{<o>}$ initial vector of hub scores
  - Compute $a^{<k>}$ and $h^{<k>}$ vectors!

Spectral Analysis of HITS- Cnt.

$$\boxed{a \leftarrow M^T h.}$$

$$\boxed{h \leftarrow M a.}$$

k=1

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

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

k=2

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

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

k=3

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

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

Spectral Analysis of HITS- Cnt.

- Unwinding the *k*-step hub-authority computation

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

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

- Do they converge to stable values?

Spectral Analysis of HITS- Cnt.

- Magnitude of hub and authority scores grow with each update.

- They only converge when we normalize them!

- In fact, it is the directions of the hub and authority vectors that are converging
  - Why?

Spectral Analysis of HITS- Cnt.

- There are normalization constants *c* and *d* so that the following vectors converge to limits as *k* goes to infinity.

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

- Let's focus on *hub* vectors (same for authority vectors)!

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

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

Spectral Analysis of HITS- Cnt.

- This

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

  - converges to **limit** ***h***$^{<*>}$, thus
    - the direction of $h^{<*>}$ at the limit should not change when multiplied with MM^T
    - Though it's length may change by a factor of c.

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

Spectral Analysis of HITS- Cnt.

- **Definition 1**: vector $\boldsymbol{v}$ is an **eigenvector** of matrix **X** if:
  - $\mathbf{X}\boldsymbol{v} = \lambda\boldsymbol{v}$
  - $\boldsymbol{v}$ an **eigenvector** of **X** and λ is its **eigenvalue**.

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

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

Spectral Analysis of HITS- Cnt.

- **Definition 2:** Any $n \times n$ **symmetric matrix** has a set of n eigenvectors that are unit vectors and mutually orthogonal
  - they form a basis for the space $\mathbb{R}^n$.

$\mathbf{X} = \mathbf{X}^T$

Spectral Analysis of HITS- Cnt.

- MM^T is symmetric!
  - Thus MM^T has n eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$
    - Let's assume that: $|\lambda_1| > |\lambda_2| >= \ldots >= |\lambda_n|$
  - Given any vector $\boldsymbol{u}$, a good way to think about $(MM^T)\boldsymbol{u}$ is to first write $\boldsymbol{u}$ as a linear combination of (MM^T)'s eigenvectors!
    - $h^{<k>} = (MM^T)^k h^{<o>}$

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

    - $h^{<k>} = (MM^T)^k (q_1 \boldsymbol{v}_1 + \ldots + q_n \boldsymbol{v}_n) =$
    - $h^{<k>} = q_1 (MM^T)^k \boldsymbol{v}_1 + \ldots + q_n (MM^T)^k \boldsymbol{v}_n =$
    - $h^{<k>} = q_1 (\lambda_1)^k \boldsymbol{v}_1 + \ldots + q_n (\lambda_n)^k \boldsymbol{v}_n$

Spectral Analysis of HITS- Cnt.

- $h^{<k>} = q_1 (\lambda_1)^k \boldsymbol{v}_1 + q_2 (\lambda_2)^k \boldsymbol{v}_2 + \ldots + q_n (\lambda_n)^k \boldsymbol{v}_n$
  - $|\lambda_1| > |\lambda_2| >= \ldots >= |\lambda_n|$

- $h^{<k>} / (\lambda_1)^k = q_1 \boldsymbol{v}_1 + q_2 (\lambda_2 / \lambda_1)^k \boldsymbol{v}_2 + \ldots + q_n (\lambda_n / \lambda_1)^k \boldsymbol{v}_n$
  - What does happen if k go to infinity?
    - every term except the first goes to 0!
    - Therefore, $h^{<k>} / (\lambda_1)^k$ converges to $q_1 \boldsymbol{v}_1$
- Remaining steps:
  - Relaxing $|\lambda_1| > |\lambda_2| >= \ldots >= |\lambda_n|$
  - See book: pages 372-374

HITS- Recap

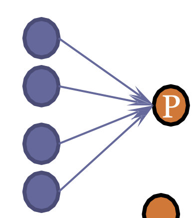
Algorithm

1. Set hub and authority scores to 1.
2. Choose a number of steps *k*.
3. Perform *k* hub-authority updates:
   1. Apply AUR to the current set of scores.
   2. Then apply HUR to the resulting scores.
4. Normalize authority and hub scores

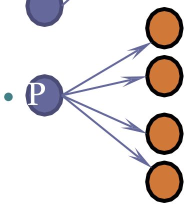
$$a^{\langle k\rangle} = (M^T M)^{k-1} M^T h^{\langle 0\rangle}$$

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

- AUR: Authority score of page p:
  - sum of the hub scores of pages that point to p.



- HUR: Hub Score of page p:
  - sum of the authority scores of pages that p points to.



Lecture Topics

- HITS
- Spectral Analysis of HITS
- **Page Rank**
- Spectral Analysis of Page Rank

Page Rank

- A page is important if it is linked / endorsed by other important pages (**iterative process**)
  - dominant mode of endorsement among
    - academic or governmental pages,
    - bloggers,
    - scientific literature, or even
    - personal pages!

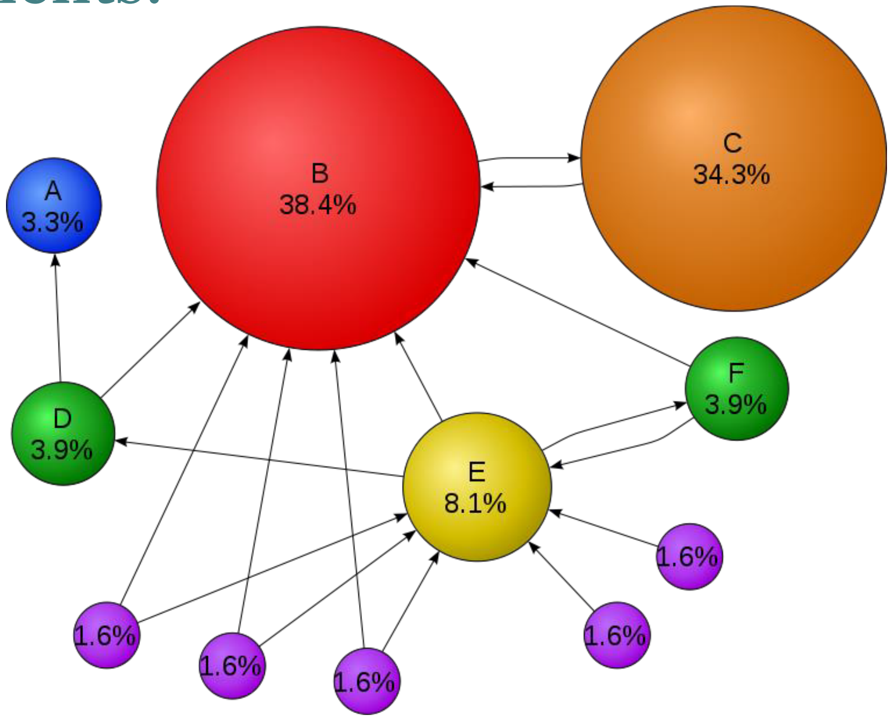
- Each node has one score, **PageRank score**!

- Votes / Endorsements pass directly from one page to another (across outgoing links)!

Named after Google co-founder Larry Page.

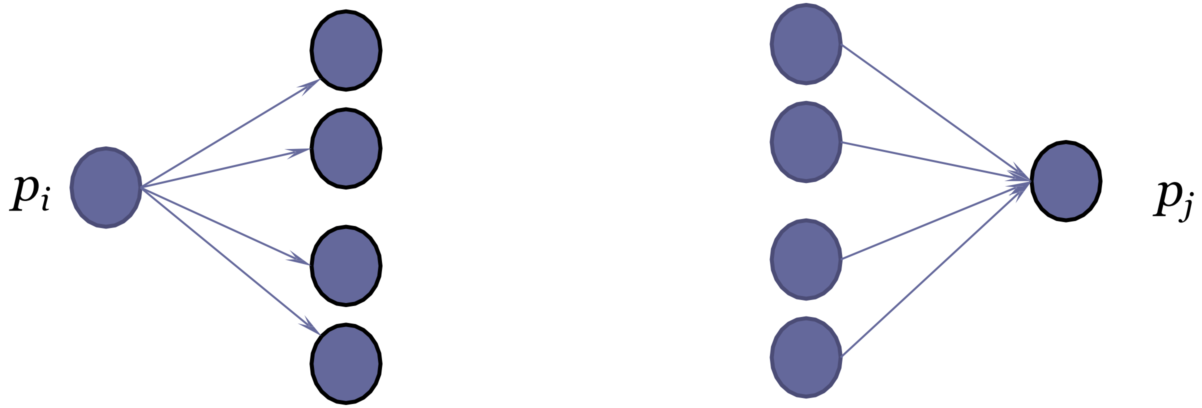
Page Rank- Cnt.

- The weight of a node's endorsement:
  - Its current PageRank score.
  - Nodes that are currently viewed as more important make stronger endorsements.



Page Rank- Cnt.

- PageRank Update Rule:
  1. Each page divides its current PageRank equally across its out-going links
     - passes **equal shares** to the pages it points to.
     - If a page has **no out-going** links, it passes all its current PageRank to itself.
  2. Each page updates its new PageRank to be the sum of the shares it receives.



Page Rank- Cnt.

Algorithm

1. Set initial PageRank of all nodes to 1/n.
2. Perform *k* updates to the PageRank values:
    1. Apply PageRank Update Rule

Page Rank- Cnt.

- PageRank intuitive view:
  - "*fluid*" that circulates through the network
  - passing across edges, and
  - pooling at the nodes that are the most important!

Page Rank- Cnt.

- Sum of PageRank values in the network?

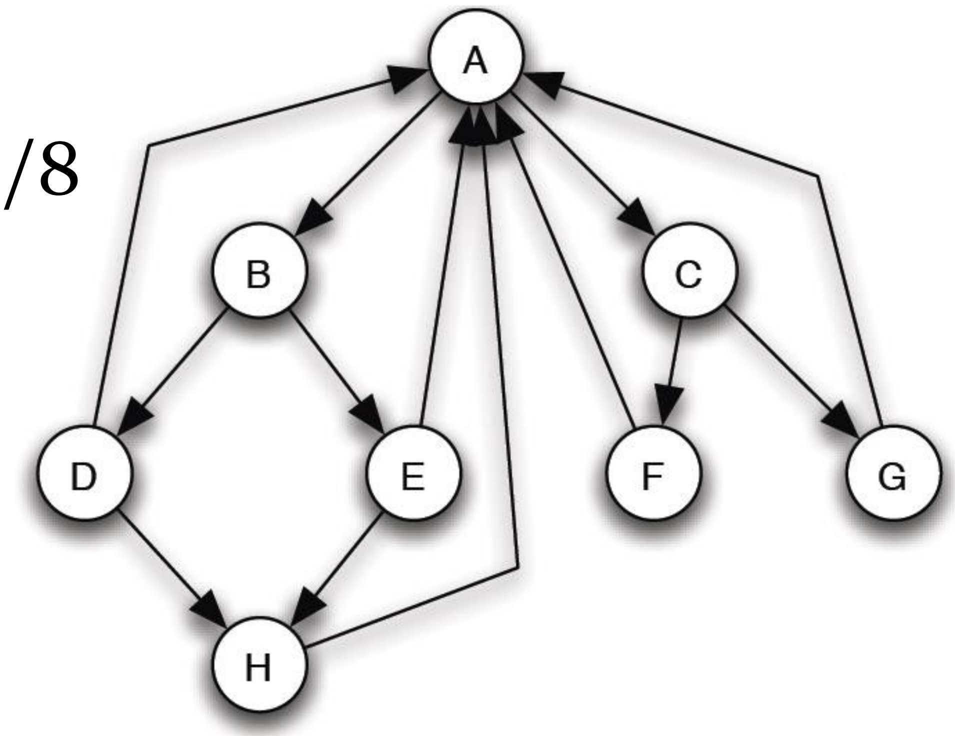
Page Rank- Cnt.

- Sum of PageRank values in the network?
  - **Remains constant** as PageRank is never created nor destroyed!
    - just moved around from one node to another.
  - Each page takes its PageRank, divides it up, and passes it along links

- We don’t need to normalize values anymore!
  - In contrast to HITS!

Page Rank- Cnt.

- n=8
- Initially p_i=1/8 for all nodes

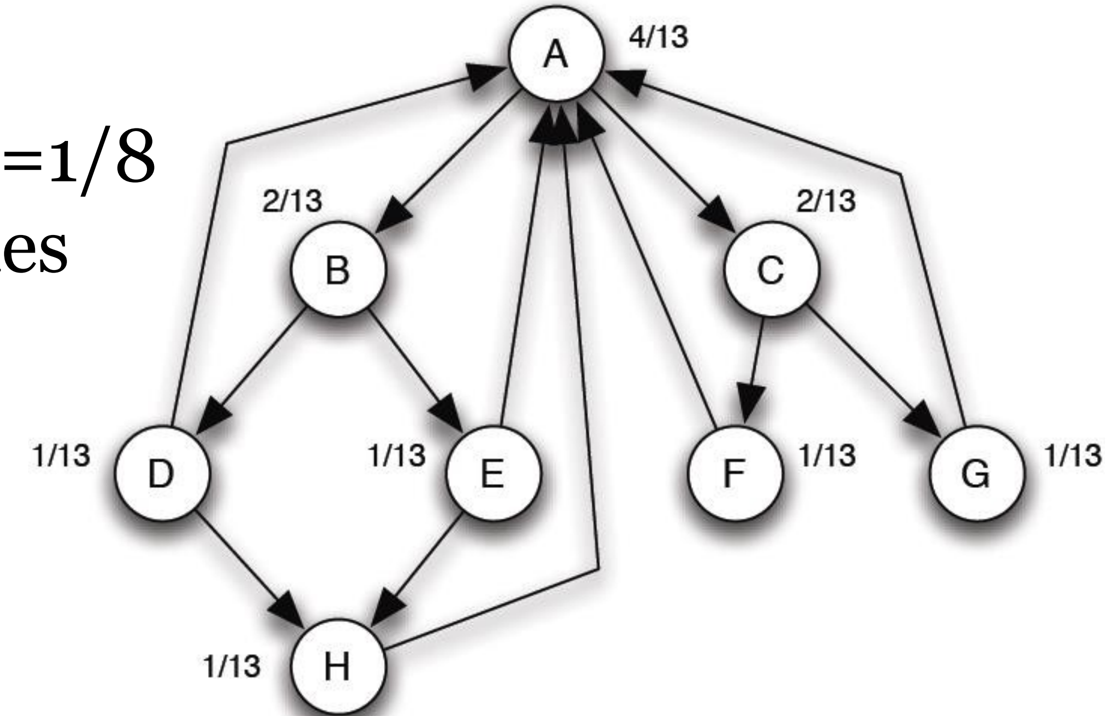


Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

A acquires a lot of PageRank
B and C benefit in the next step.
B and C are more important than D, E, F, G, and H

Page Rank- Cnt.

- n=8
- Initially p_i=1/8 for all nodes

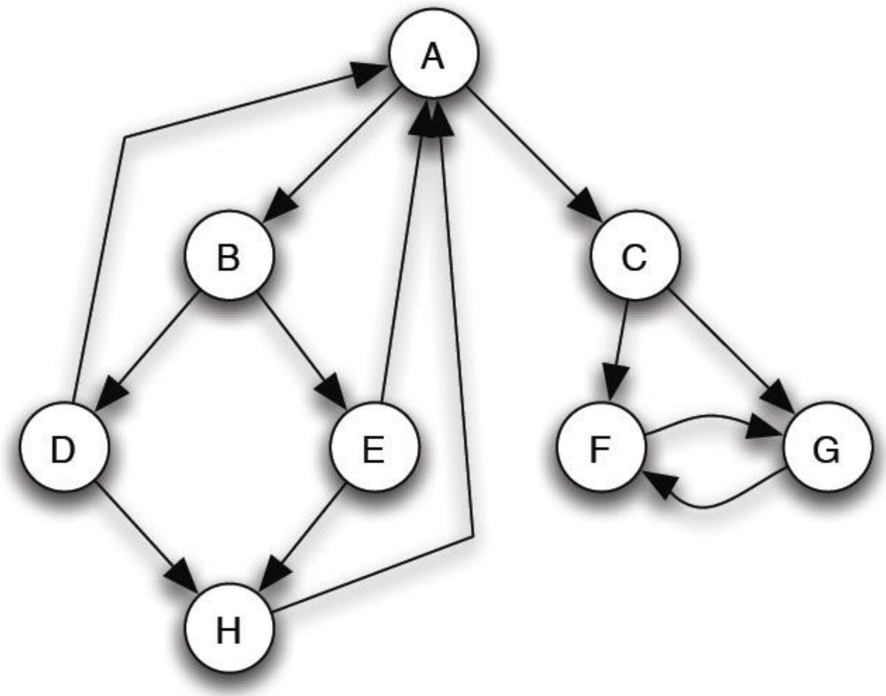


<u>Equilibrium Values of PageRank</u>

When we reach the limiting PageRank values:
1. PageRank values sum to 1, and
2. If we apply the PageRank Update Rule, the values at every node remains the same
   - values regenerate themselves exactly when they are updated.

Page Rank- Cnt.

- Issue with page rank algorithm?



Page Rank- Cnt.

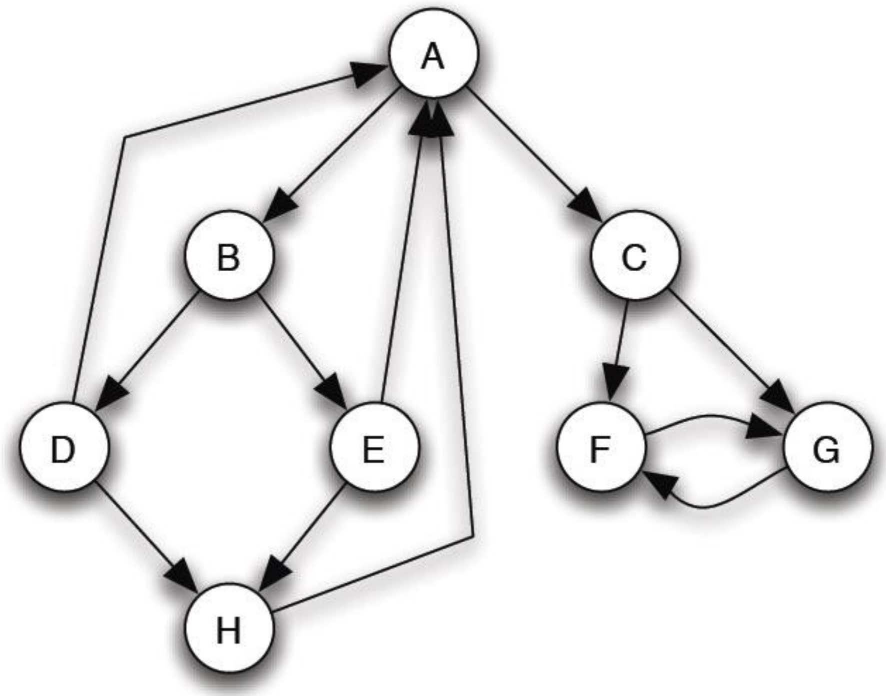
- Issue with page rank algorithm?
  - "Wrong" nodes can end up with **all the PageRank** in the network!

F and G point to each other!

PageRank that flows from C to F and G can never circulate back into the rest of the network

For large *k*, PageRank values converge to 1/2 for each of F and G, and 0 for all other nodes.

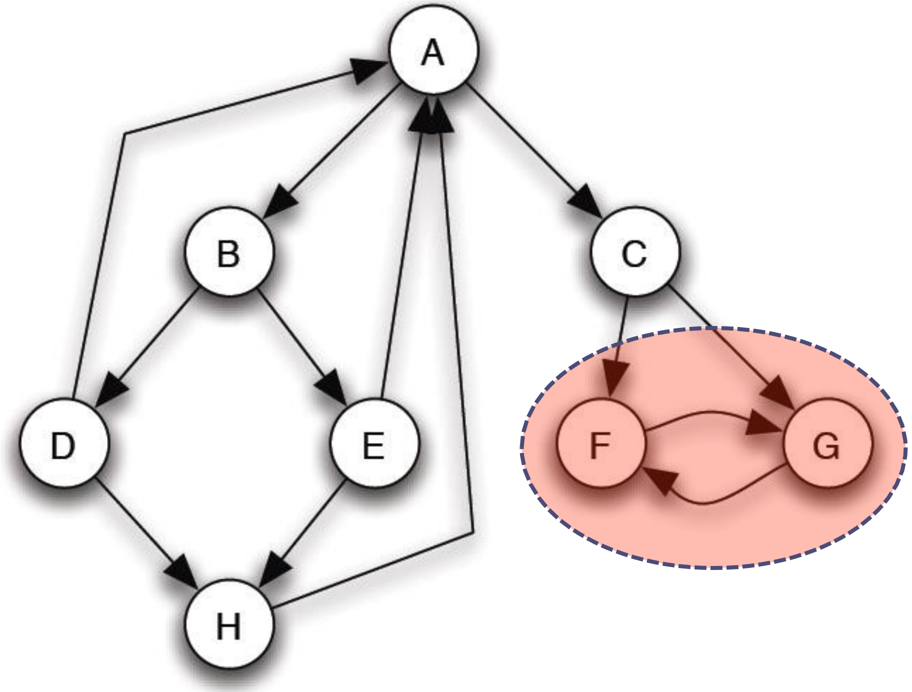
The links out of C function as a kind of "**slow leak**" that eventually causes all the PageRank to end up at F and G.



Page Rank- Cnt.

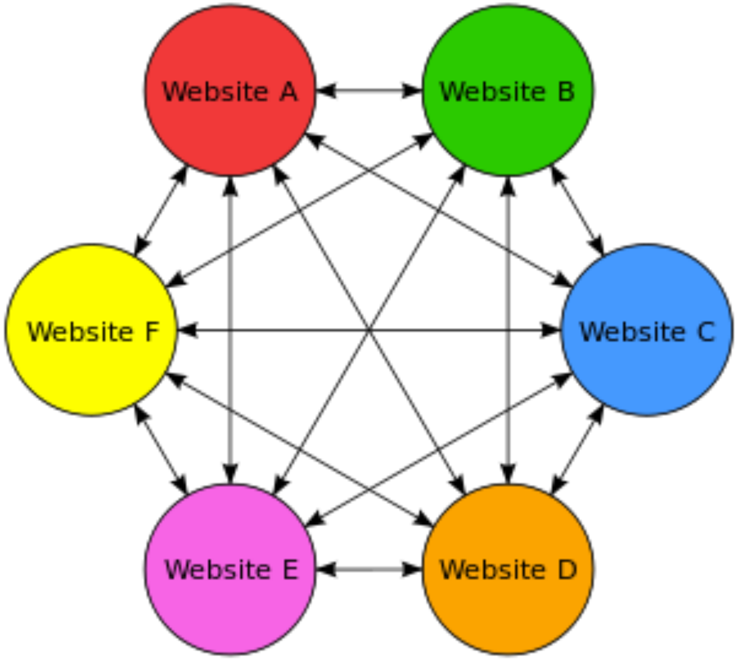
- Issue with page rank algorithm?
  - “Wrong” nodes can end up with **all the PageRank** in the network!

As long as there are small sets of nodes that can be reached from the rest of the graph, but have no paths back, then PageRank will build up there!



Page Rank- Cnt.

- Link farms
  - Pages heavily linked to each other
  - Created by automated programs
  - Fooling search engines



Page Rank- Cnt.

- Scaled PageRank Update Rule:
  - Pick a scaling factor $0 < s < 1$
  - Apply the PageRank Update Rule as before.
  - Then scale down all PageRank values by a factor of s.
    - The total PageRank in the network?
      - shrinks from 1 to s.
  - Divide residual (1–s) PageRank equally over nodes
    - giving (1- s) /n to each node.

Common values for *s* are in the range of 0.8 to 0.9!

The above rule follows from the "fluid" intuition for PageRank

- Why all the water on earth doesn't inexorably run downhill and reside exclusively at the lowest points?
- There's a counter-balancing process at work:
- Water also evaporates and gets rained back down at higher elevations!

Page Rank- Cnt.

- Repeated application of the Scaled PageRank Update Rule **converges** to a set of limiting PageRank values as k goes to infinity!

- Do the resulting values depend on the choice of s?
  - Yes, different update rules for different values of s.

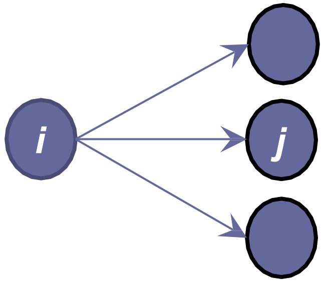
Lecture Topics

- HITS
- Spectral Analysis of HITS
- Page Rank
- **Spectral Analysis of Page Rank**

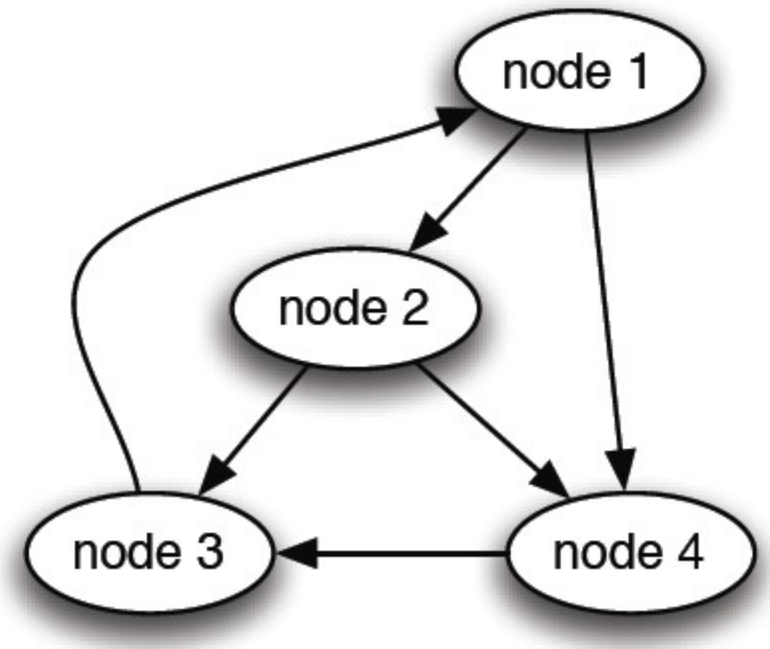
Spectral Analysis of PageRank

<u>PageRank Update Rule</u>

- Let $\mathbf{N}_{ij}$ to be the portion of i's PageRank that circulate to j in one update step:
  - $\mathbf{N}_{ij} = 0$ if i doesn't link to j
  - $\mathbf{N}_{ij} = 1/l_i$ Otherwise
    - Where l_i is out-degree of i
  - If i has no outgoing links, then we define $\mathbf{N}_{ii} = 1$
    - A node with no outgoing links passes all its PageRank to itself



Spectral Analysis of PageRank- Cnt.



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

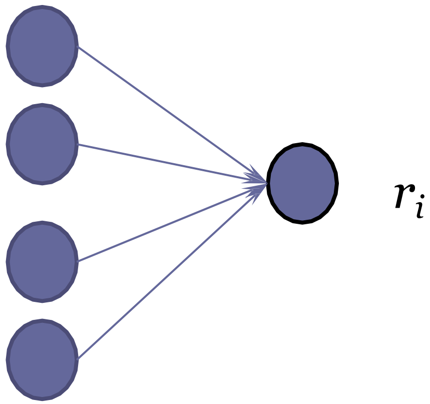
Figure 14.13: The flow of PageRank under the Basic PageRank Update Rule can be represented using a matrix N derived from the adjacency matrix M: the entry N_{ij} specifies the portion of i's PageRank that should be passed to j in one update step.

Spectral Analysis of PageRank- Cnt.

- $r \in n \times 1$
  - vector representing PageRanks of all n nodes
- PageRank Update Rule:

$$r_i \leftarrow N_{1i}r_1 + N_{2i}r_2 + \cdots + N_{ni}r_n.$$

$$r \leftarrow N^T r.$$



Spectral Analysis of PageRank- Cnt.

Scaled PageRank Update Rule

- Let $\tilde{\mathbf{N}}_{ij}$ to be the portion of *i*'s PageRank that circulate to *j* in one update step:
  - The updated PageRank is scaled down by a factor of *s*, and the residual (1 – s) units are divided equally over all nodes

$$r_i \leftarrow \tilde{N}_{1i}r_1 + \tilde{N}_{2i}r_2 + \cdots + \tilde{N}_{ni}r_n.$$

$$r \leftarrow \tilde{N}^T r.$$

Spectral Analysis of PageRank- Cnt.

- Start from an initial PageRank vector $r^{<0>}$ and produce a sequence of vectors $r^{<1>}$, $r^{<2>}$, . . . each is obtained from the preceding one via multiplication by $\tilde{\mathbf{N}}^{T}$.
  - Unwind this process:

$$r^{\langle k \rangle} = (\tilde{N}^T)^k r^{\langle 0 \rangle}.$$

- The Scaled PageRank Update Rule converges to a limiting vector $r^{<*>}$!

$$\tilde{N}^T r^{\langle * \rangle} = r^{\langle * \rangle}$$

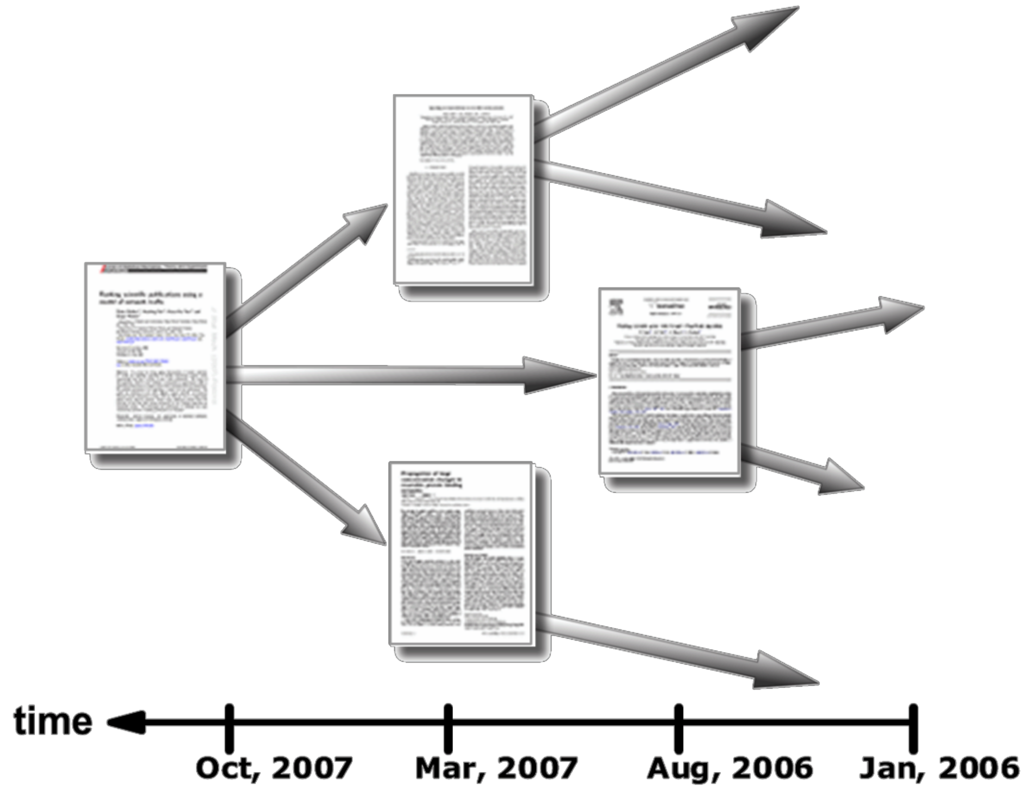
  - $r^{<*>}$ should be an eigenvector of $\tilde{\mathbf{N}}^{T}$ with eigenvalue of 1.
- See book: page 376.

Applications

- Impact Factor of Scientific Journals

Impact Factor for a scientific journal: The average number of citations received by papers published in the given journal over the past two years.

In-links indicate **collective attention** that the scientific community pays to papers published in the journal.

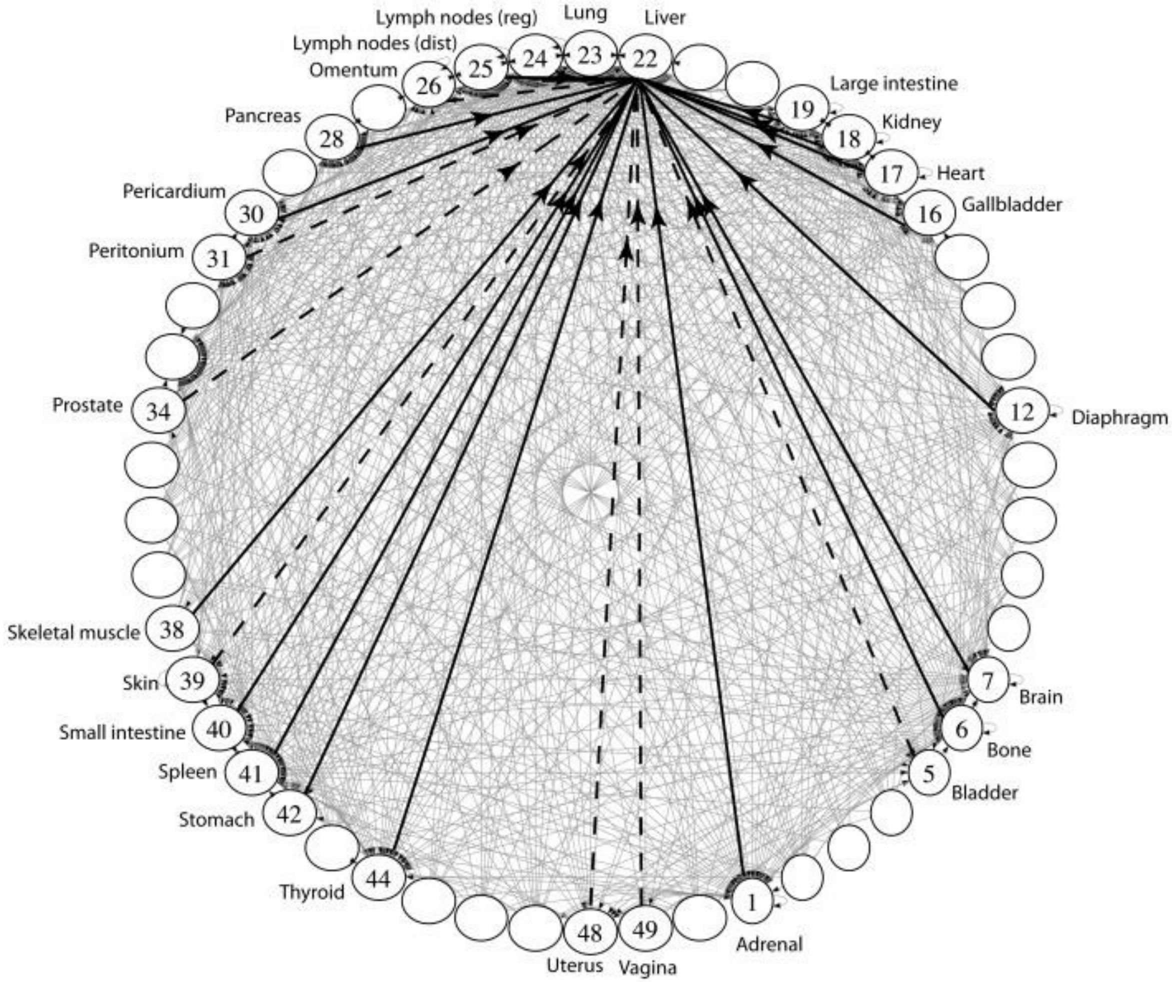


Applications- Cnt.

- Fighting Lung Cancer Using PageRank

metastatic lung cancer does not progress in a single direction from primary tumor site to distant locations, **which has been the traditional medical view**. Instead … cancer cell movement around the body likely occurs in more than one direction at a time.

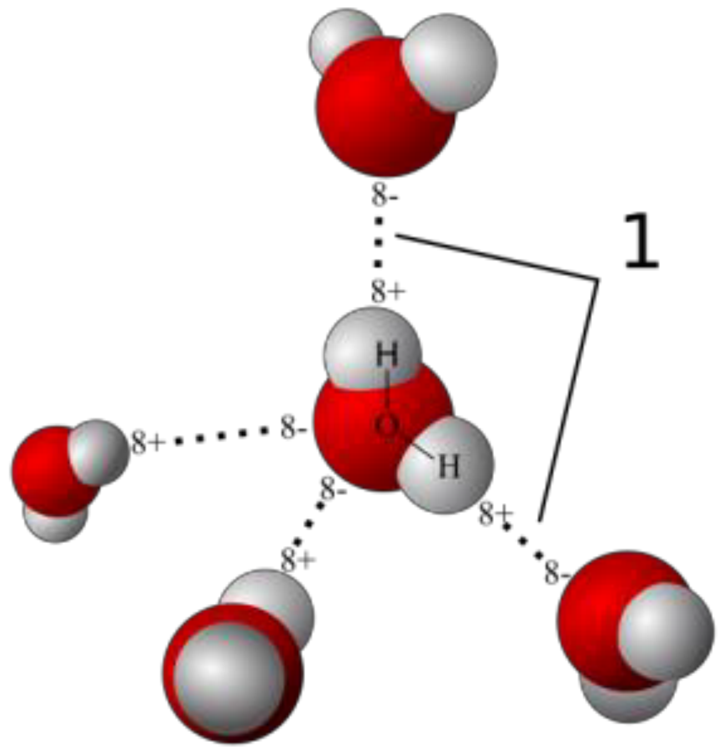
certain organs tend to spread cancer cells more aggressively, while others tend to act as sponges for cancer cells!



Applications- Cnt.

- Applying PageRank to the Molecular Universe

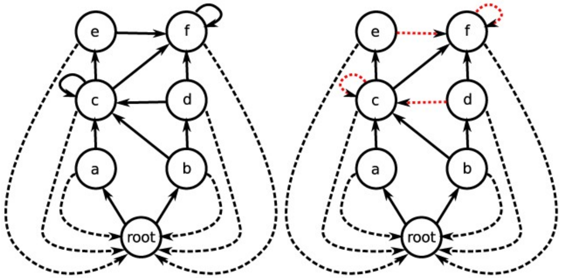
Because the **PageRank of a molecule affects how it will act in a chemical reaction** — and water is involved in almost every biological process. By understanding how a network of trillions of molecules interact, scientists can produce much more accurate models of chemical reactions.



Applications- Cnt.

- Google trick tracks extinctions

Google's algorithm for ranking web pages can be adapted to **determine which species are critical for sustaining ecosystems**.



Modification of food webs from ecological considerations to satisfy the two constraints required for application of the algorithm.

Reading

- Ch.14 Link Analysis and Web search [NCM]
- Ch.05 Link Analysis [MMD]
- The anatomy of the Facebook social graph. Ugander, J., et al. arXiv'11.
- Four degrees of separation . Backstrom L., et al. WebSci'12.
- The small world problem . Milgram S. Psychology Today'1967.