CS 6604: Data Mining Large Networks and Time-series
Fall 2013

CENTRALITY MEASURES

Shaikh Arifuzzaman and Md Hasanuzzaman Bhuiyan

Outline

- Part 1
  - Basic Centrality Concepts
    - Degree Centrality
    - Betweenness Centrality
    - Closeness Centrality
    - Eigenvector Centrality
    - Centralization
- Part 2
  - Part 2A
    - Hub and Authorities (HITS Algorithm)
    - PageRank
  - Part 2B
    - Spectral Analysis of Hub and Authorities
    - Spectral Analysis of PageRank

PART 1

Basic centrality concepts

Centrality

- Relative importance of a node in the graph
- Which nodes are in the "center" of a graph?
  - What do you mean by "center"?
  - Definition of "center" varies by context/purpose
- "There is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is little agreement on the proper procedure for its measurement."
  - by Freeman, 1979

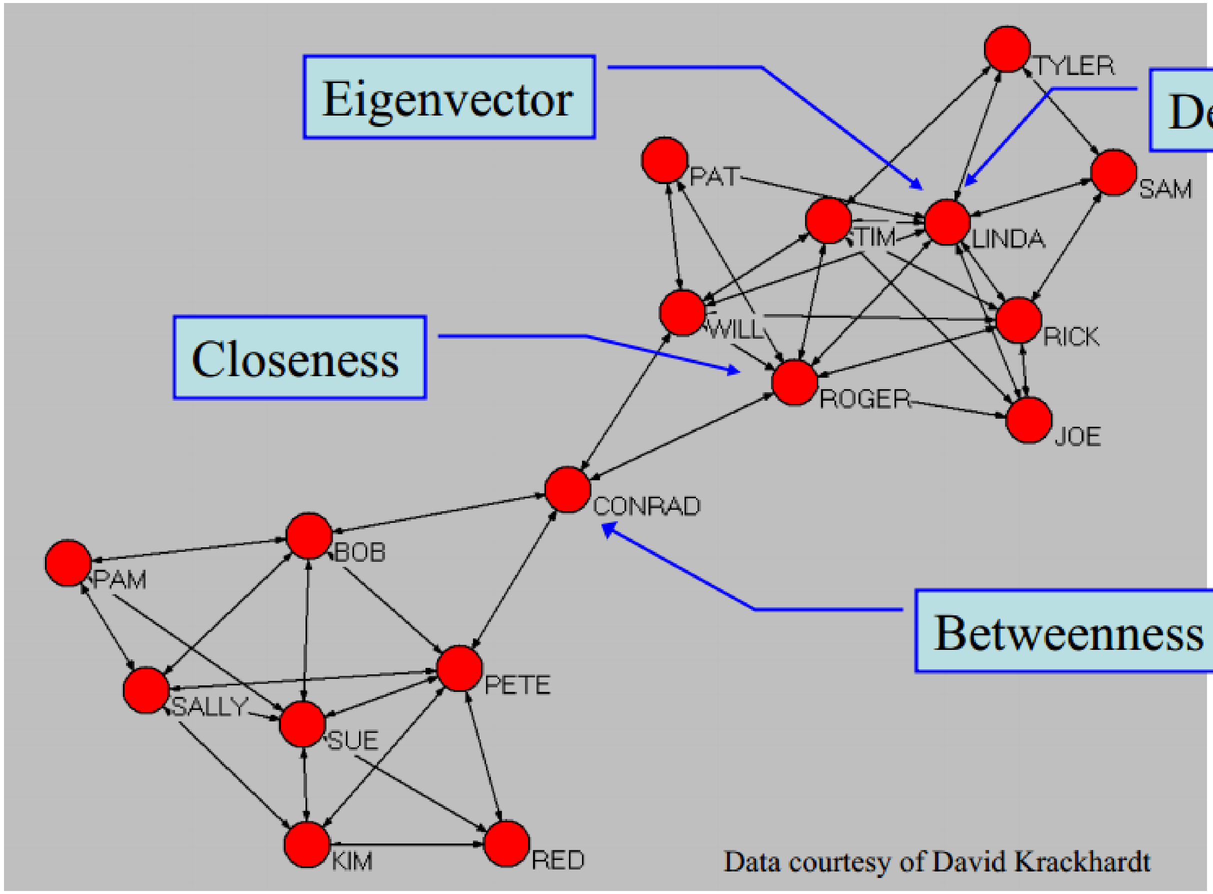
Centrality

- Real valued function on the nodes of a graph
- Structural index
- Applications:
  - How influential a person is in a social network?
  - How well used a road is in a transportation network?
  - How important a web page is?
  - How important a room is in a buildling?

Centrality Measures

- Different measures of centrality:
  - Degree centrality
  - Betweenness centrality
  - Closeness centrality
  - Eigenvector centrality

Example [Borgatti, 2005]



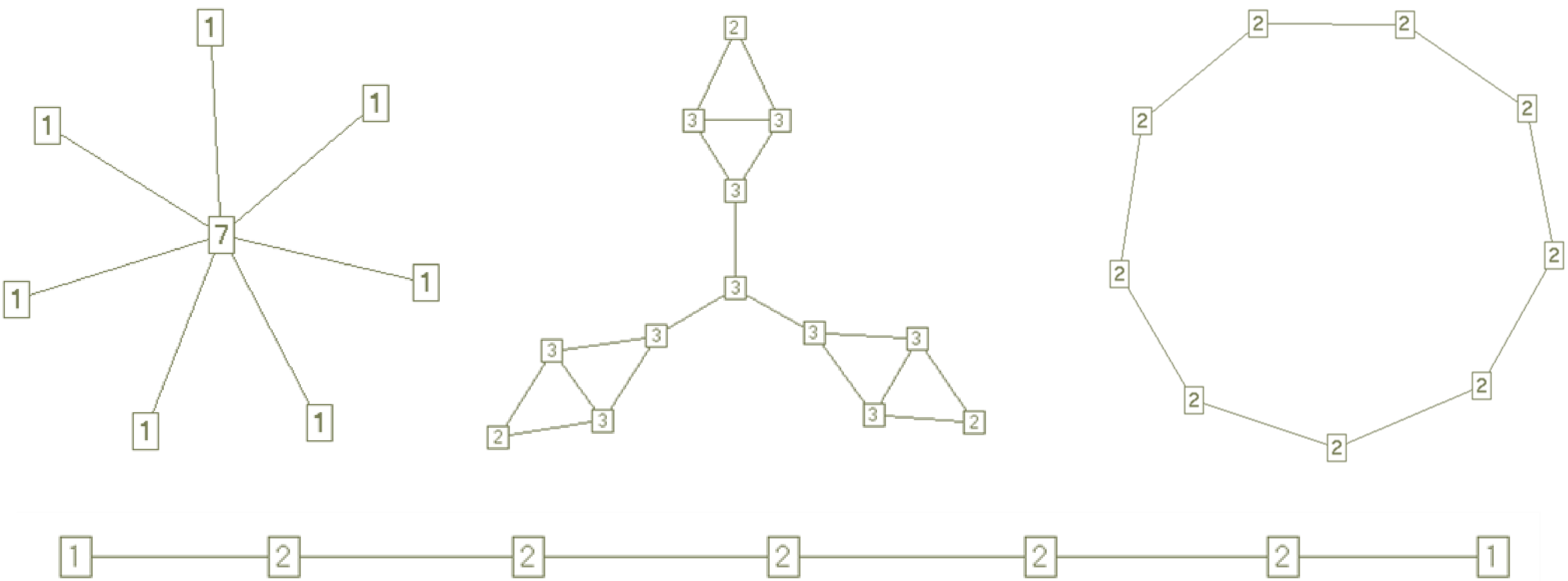
Degree Centrality

Degree Centrality

- Most intuitive notion of centrality
- Node with the highest degree is most important
- Index of exposure to what is flowing through the network
  - Gossip network: central actor more likely to hear a gossip
- Normalized degree centrality
  - Divide by max. possible degree (n-1)

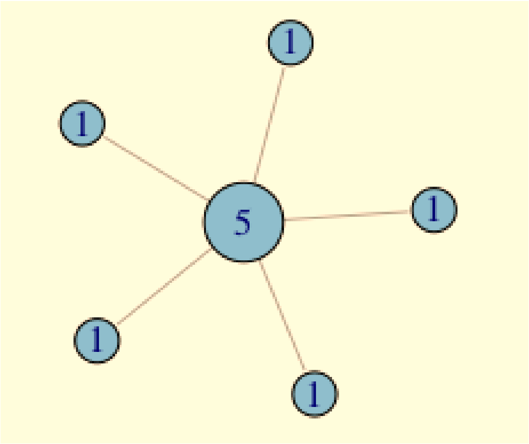
Degree Centrality

- Example:



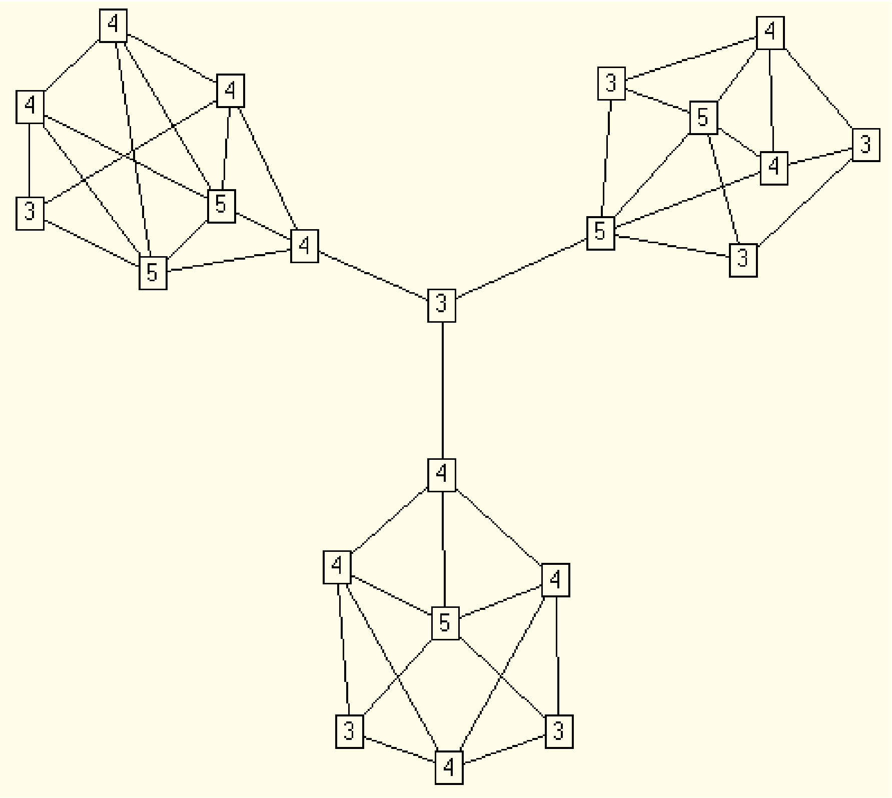
Degree Centrality

- When to use?
  - Whom to ask for favor?
  - People you can talk to



Degree Centrality

- Can be deceiving
  - Why?
    - Local measure



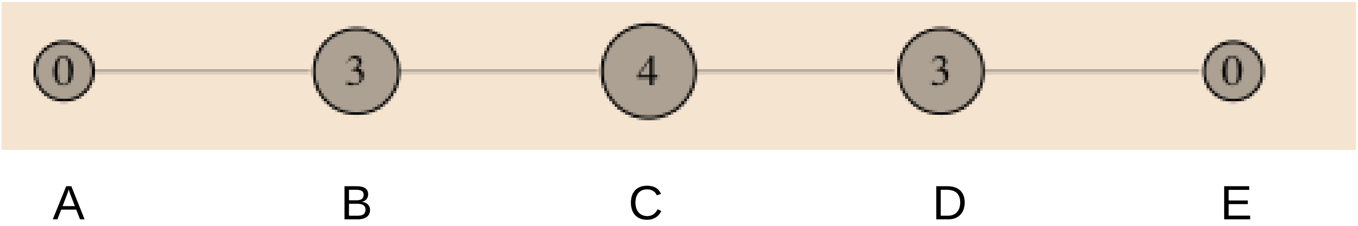
Betweenness Centrality

Betweenness Centrality

- BC of a node u is the ratio of the shortest paths between all other nodes, that pass through node u
- Quantifies the control of a node on the communication between other nodes
- First introduced by Freeman
- $C_B(u) = \sum_{s \neq v \neq t} \frac{\delta_{st}(u)}{\delta_{st}}$
  - s = source
  - t = destination
  - δ_{st} = number of shortest paths between (s, t)
  - $\delta_{st}(u)$ = number of shortest paths between (s, t) that pass through u

Betweenness Centrality

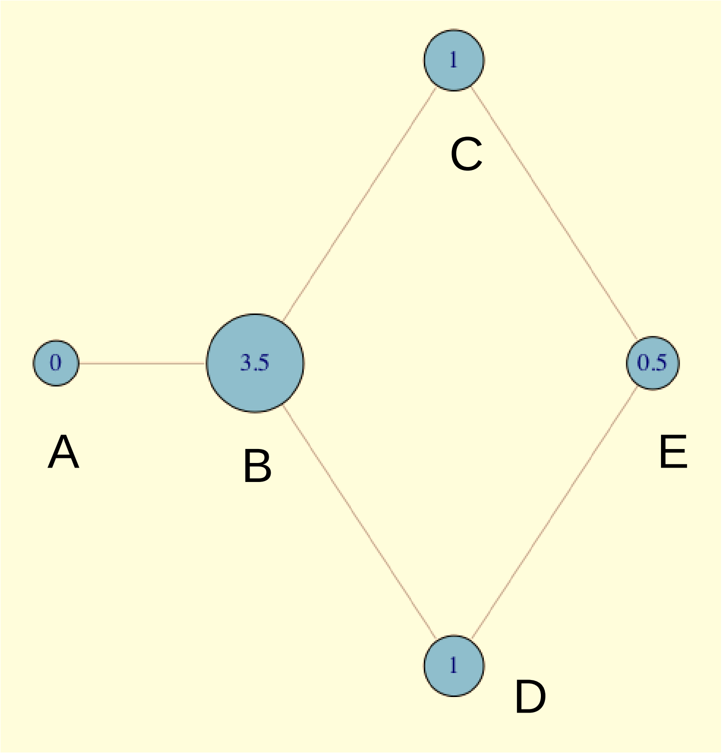
- Example:



- A lies between no two other vertices
- B lies between A and 3 other vertices: C, D, and E
- C lies between 4 pairs of vertices $(A, D), (A, E), (B, D), (B, E)$

Betweenness Centrality

- More Example:
- why do C and D each have betweenness 1?
- They are both on shortest paths for pairs (A,E), and (B,E), and so must share credit:
  - $\frac{1}{2}+\frac{1}{2} = 1$
- Can you figure out why B has betweenness 3.5 while E has betweenness 0.5?

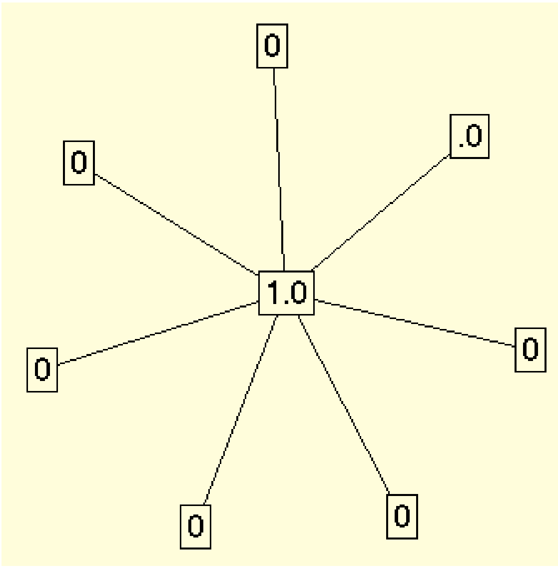


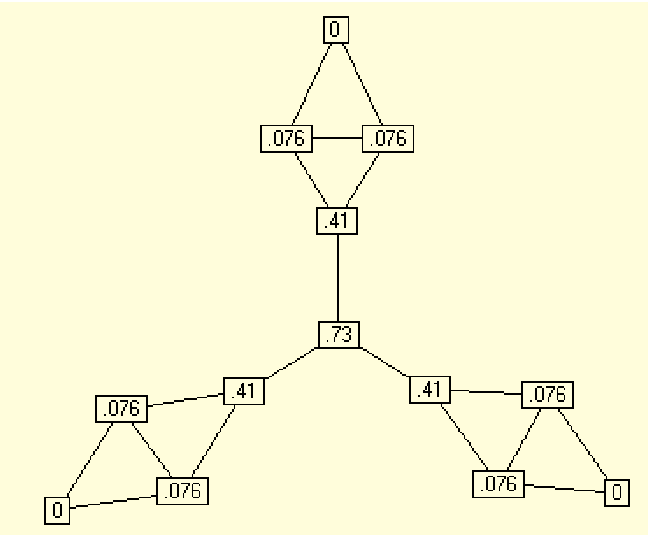
Betweenness Centrality

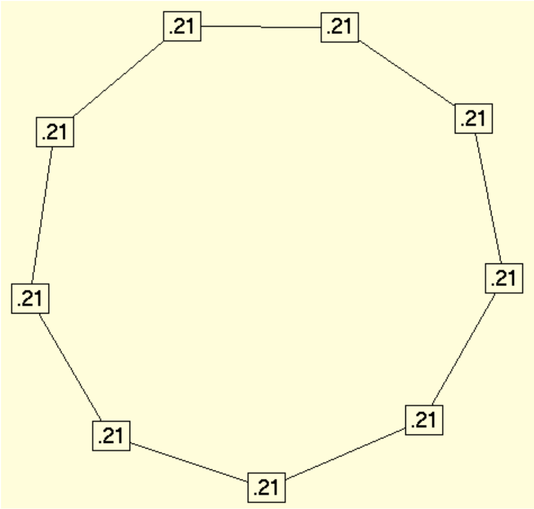
- Famous algorithm by Brandes
  - $O(mn)$ for unweighted graph
  - $O(n^2 \log n + mn)$ for weighted graph
- Edge betweenness centrality
  - Pass through that edge
- Normalize
  - Divide by $\binom{n-1}{2}$ for undirected graph
    - Number of pairs of nodes excluding itself
  - Divide by $(n-1)(n-2)$ for directed graph

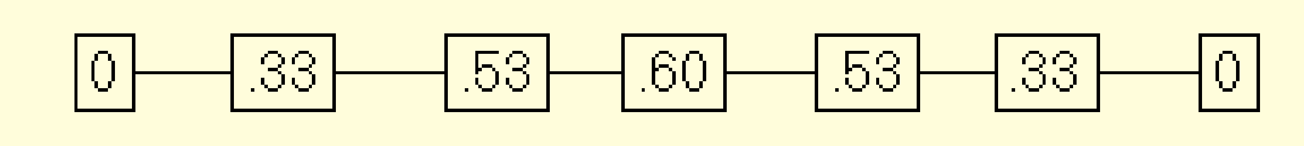
Betweenness Centrality

- Normalized example:



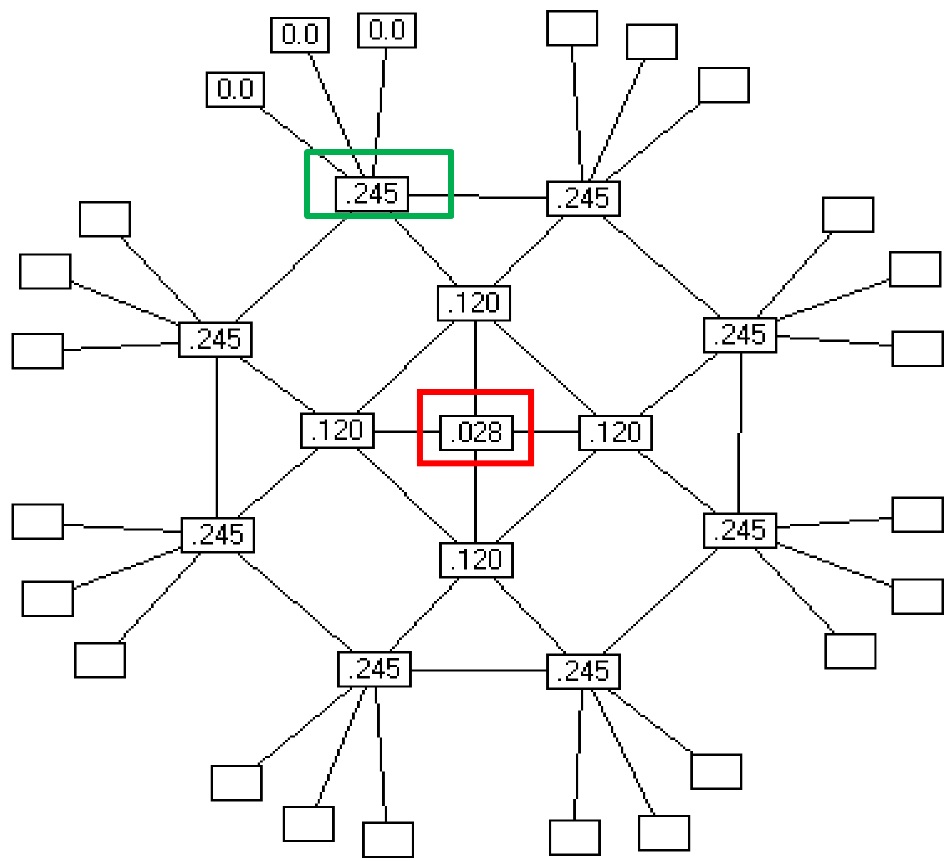






Betweenness Centrality

- Normalized example:
- Red circled node has low centrality value. Why?
- Green circled node has high value. Why?



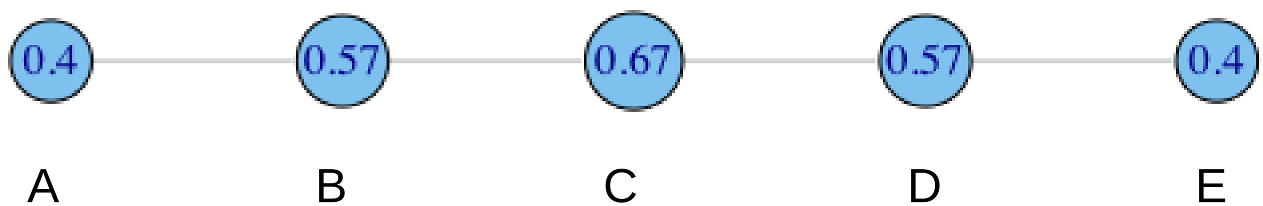
Closeness Centrality

Closeness Centrality

- A node is considered important if it is relatively close to all other nodes.
- Farness of a node is the sum of its distances to all other nodes.
- Closeness if the inverse of the farness.
- $C_C(u) = \frac{1}{\sum_{v \neq u} d(u,v)}$
- Normalized:
  - Divide by $(n-1)$

Closeness Centrality

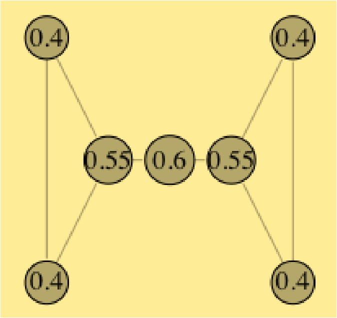
- Closeness is a measure of how long it will take to spread information from node u to all other nodes
- Normalized Example:

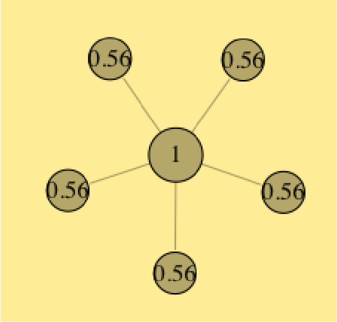


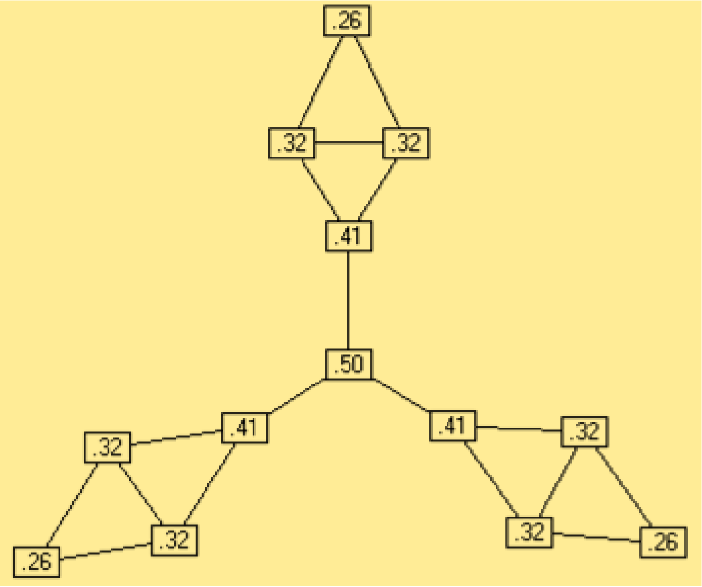
$$C_c^{'}(A) = \left[\frac{\sum_{j=1}^{N} d(A, j)}{N-1} \right]^{-1} = \left[\frac{1+2+3+4}{4} \right]^{-1} = \left[\frac{10}{4} \right]^{-1} = 0.4$$

Closeness Centrality

- More example:







Comparison

- Comparing across 3 centrality values
  - Generally, the 3 types will be positively correlated
  - When they are not, it tells you something interesting!

	Low Degree	Low Closeness	Low Betweenness
High Degree		Embedded in cluster that is far from the rest of the network	Ego's connections are redundant - communication bypasses him/her
High Closeness	Key player tied to important/active alters		Probably multiple paths in the network, ego is near many people, but so are many others
High Betweenness	Ego's few ties are crucial for network flow	Very rare cell. Would mean that ego monopolizes the ties from a small number of people to many others.	

Eigenvector Centrality

- Measure of the influence of a node in a network
- Connections to high-scoring nodes contribute more
- "An important node is connected to important neighbor"
- Google's PageRank is a variant of Eigenvector centrality
- Eigenvector centrality of v, $x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$
- $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$
- Power iteration is one of the eigenvalue algorithm

Centralization of Network

- Measure of how central its most central node is in relation to how central all the other nodes are
- How much variation in the centrality scores?
- Every centrality measure can have its own centralization measure
- Freeman's formula for centralization of degree:

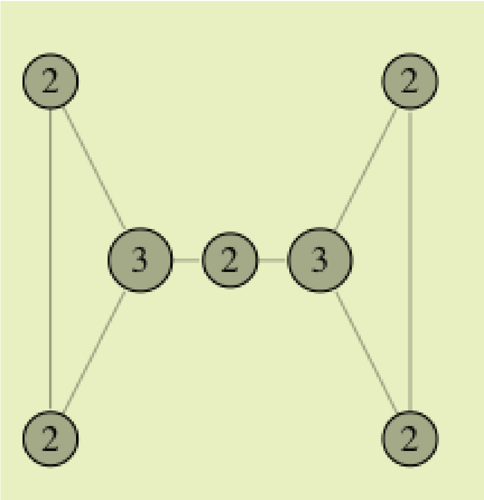
maximum value in the network

$$C_D = \frac{\sum_{i=1}^{n}[\boxed{C_D(n^*)} - C_D(i)]}{\boxed{(N-1)(N-2)}}$$

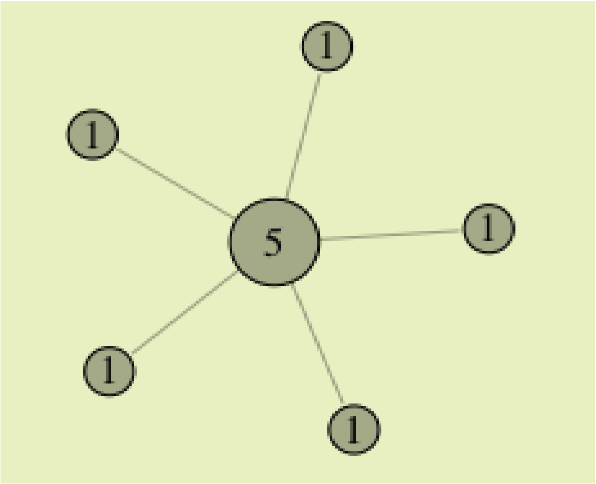
theoretically largest such sum of differences in any network of the same degree

Centralization of Network

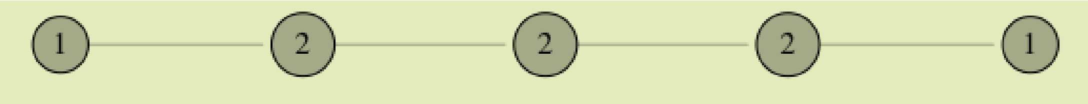
- Degree Centralization Example:



$C_D = 0.167$



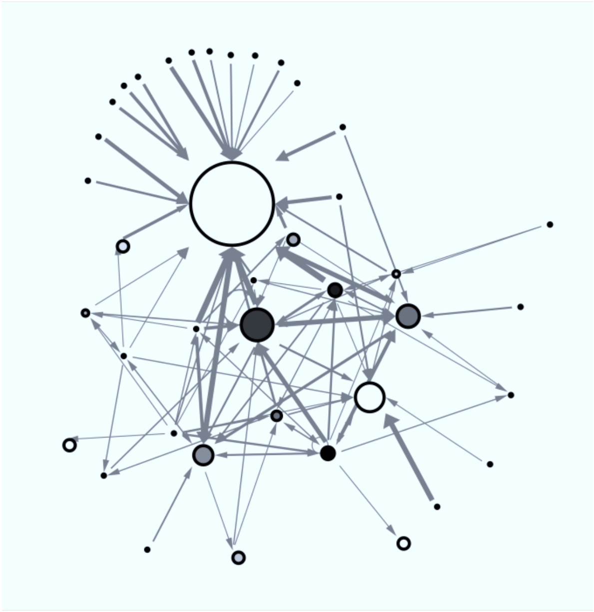
$C_D = 1.0$



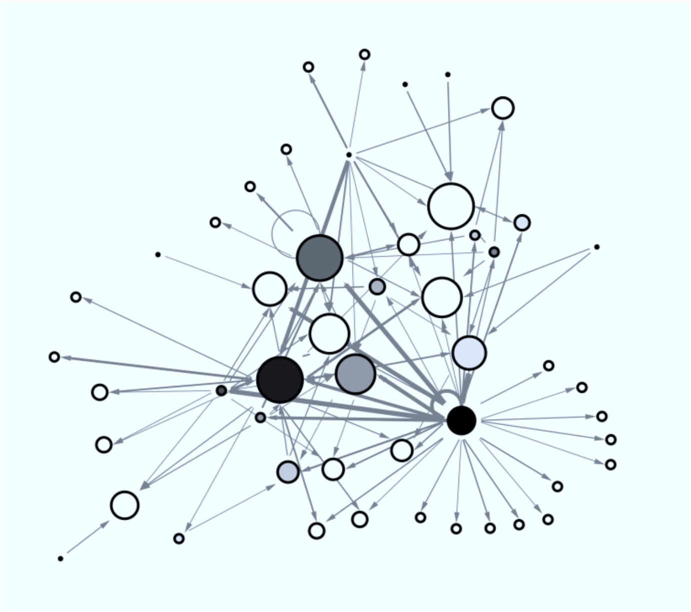
$C_D = 0.167$

Centralization of Network

- Degree Centralization Example: financial trading networks



high centralization: one node trading with many others



low centralization: trades are more evenly distributed

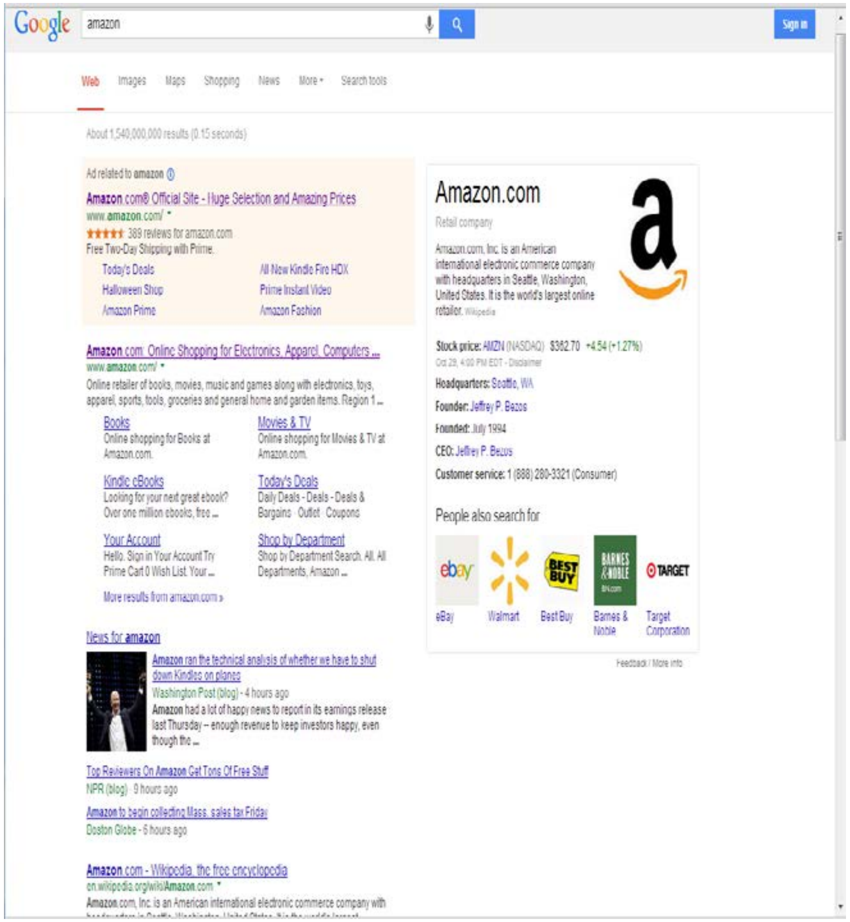
PART 2A

Hub-Authority and PageRank: Conceptual Introduction

Searching the Web

- How does Google know the "best" answers?
- How hard is the problem?
  - Synonymy
  - Polysemy
  - dynamicity

Understanding the network structure of web pages is crucial

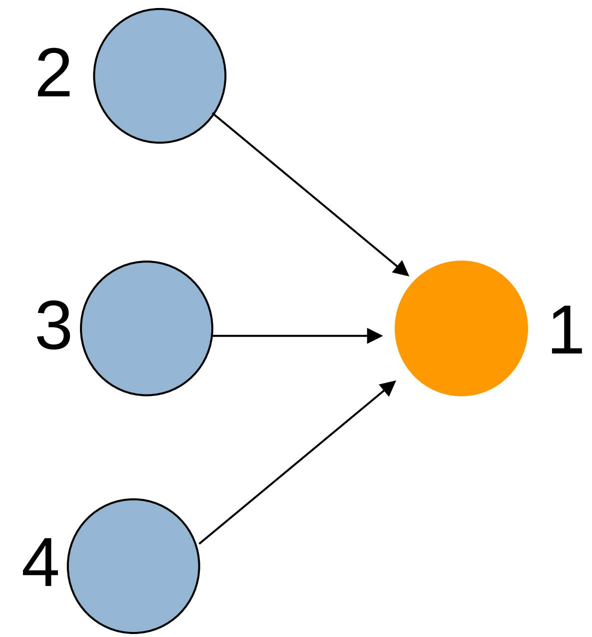


Link Analysis

- In this hyperlinked network of webpages, which pages are most popular/important?
  - More in-links?
  - More out-links?
  - Combinations?

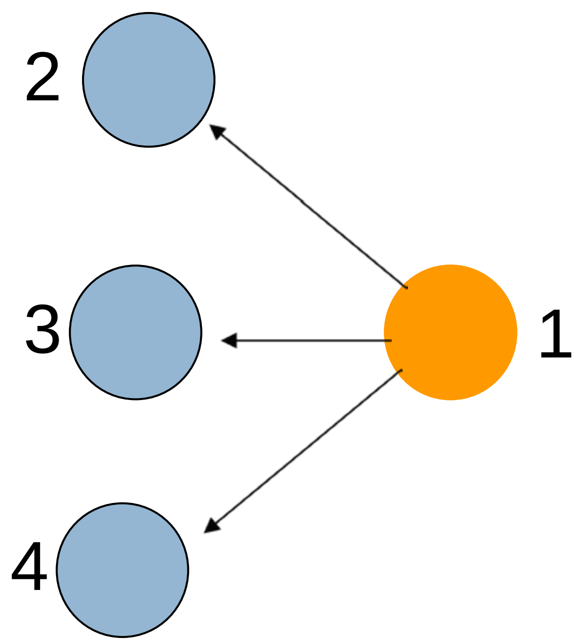
Voting by in-links

- How to rank pages
  - From in-links?
- Intuition:
  - Implicit endorsement
  - Single vs aggregate endorsement
  - Page referred by most preferred



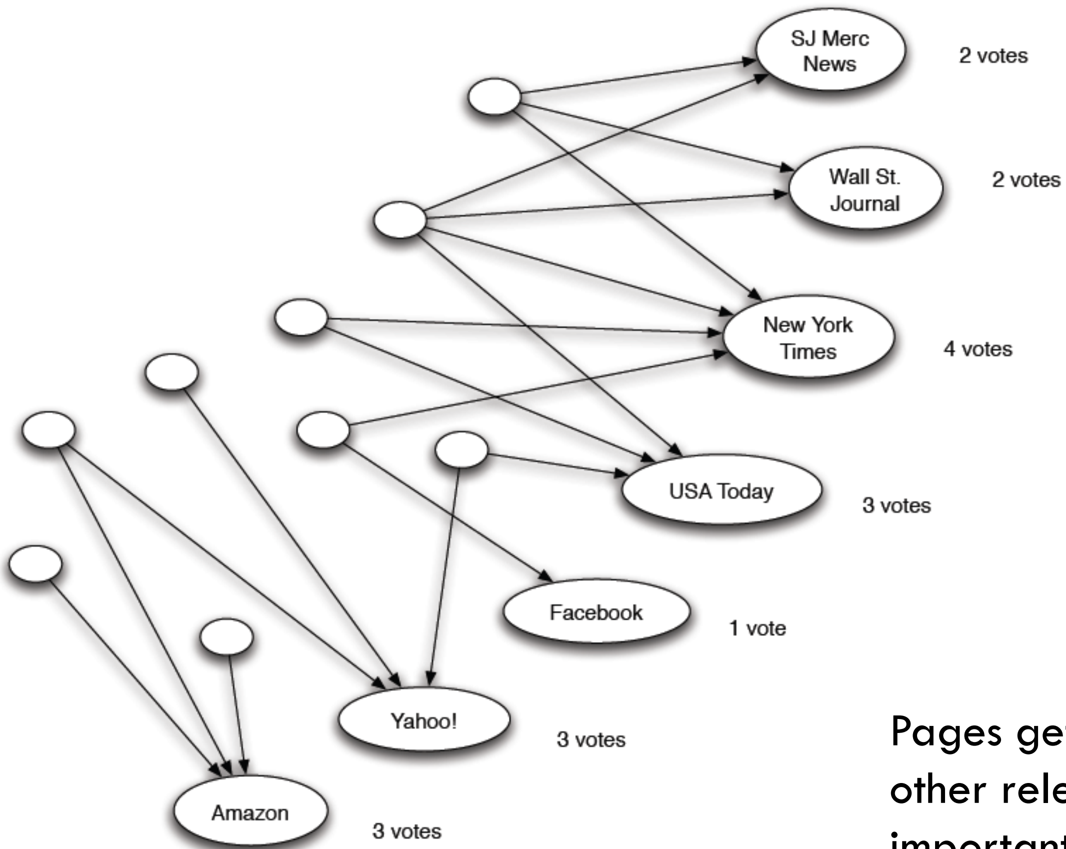
How about out-links

- Any implication of out-links?



An example [Kleinberg]

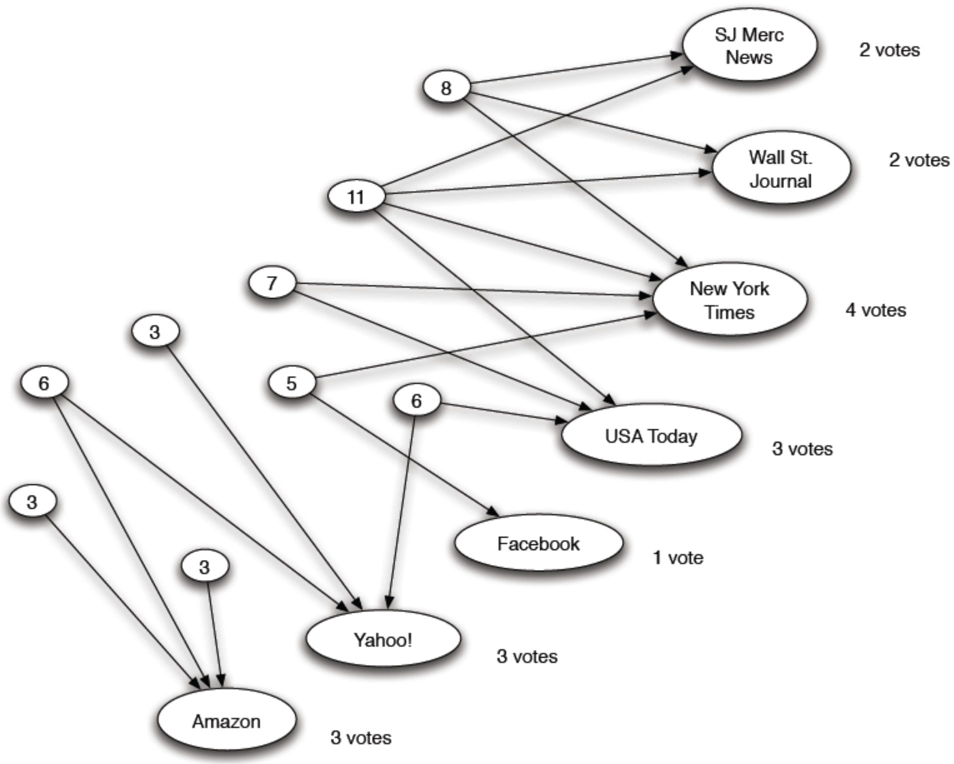
In-links to pages for the query newspaper



Pages getting higher in-links from other relevant pages are important

An example [Kleinberg] contd.

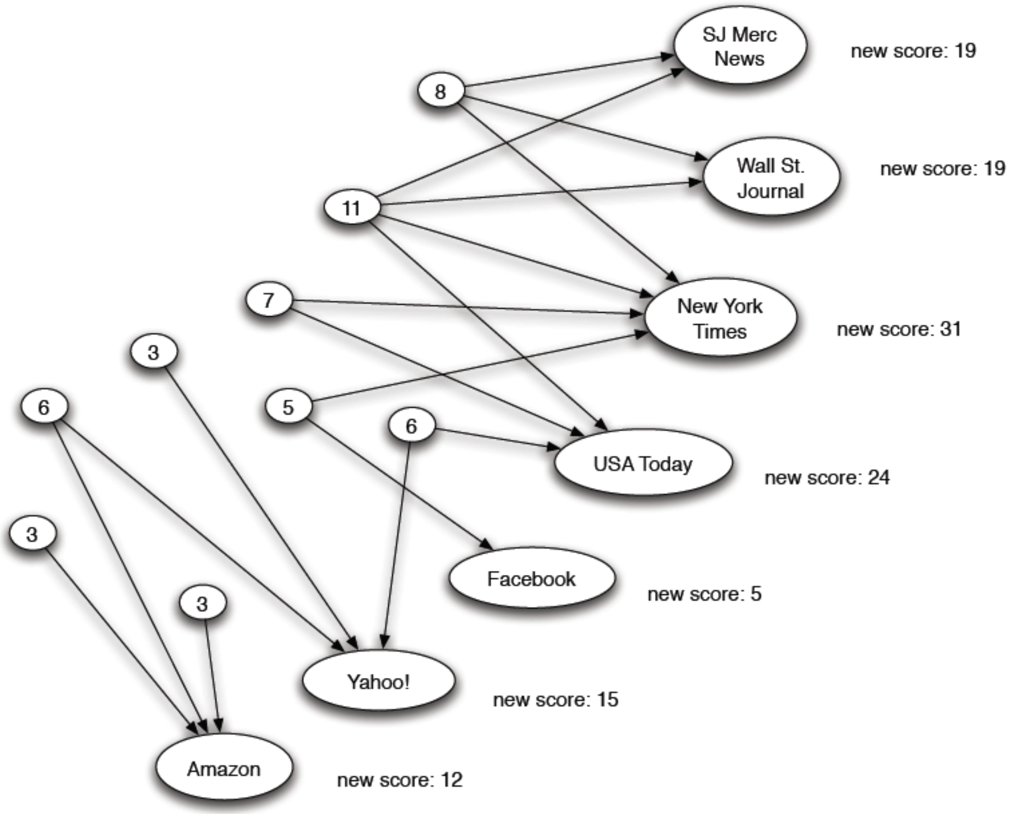
Good lists: some pages compile lists of relevant resources



Pages listing higher number of relevant resources should score higher as lists

An example [Kleinberg] contd.

Updated score: some of scores of all lists that point to it



Where does it head to?

- Principle of repeated improvement

Hub-Authority (HITS Algorithm)

- Authority: highly endorsed answers to queries
- Hub: high value lists for the query

Quality of hubs to refine estimate of the quality of the authorities

- Authority update rule
- Hub update rule

- Recursive dependency:

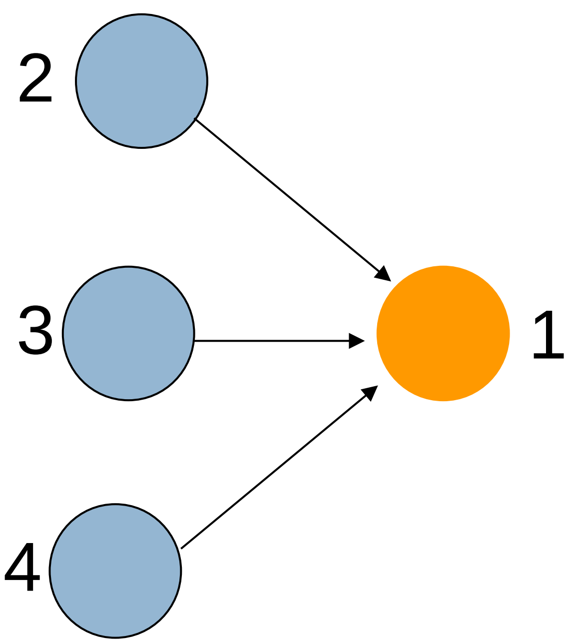
$$a(v) \leftarrow \Sigma_{w \epsilon parent[v]} h(w)$$

$$h(v) \leftarrow \Sigma_{w \epsilon children[v]} a(w)$$

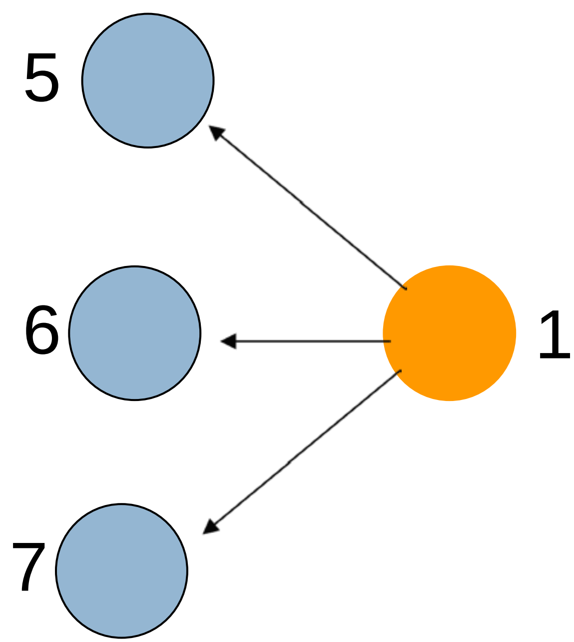
Hub-Authority (HITS Algorithm)

- Authority: highly endorsed answers to queries
- Hub: high value lists for the query

$$a(1) = h(2) + h(3) + h(4)$$



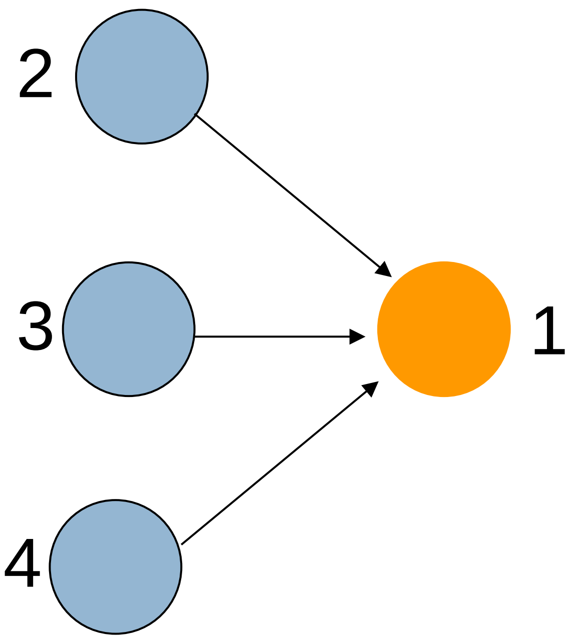
$$h(1) = a(5) + a(6) + a(7)$$



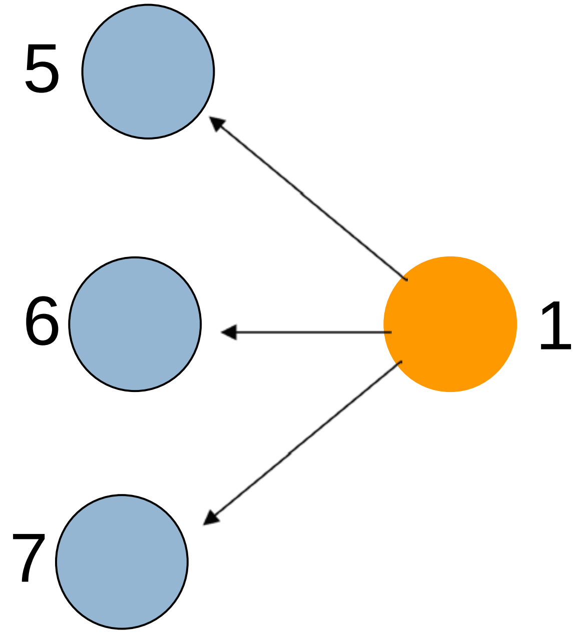
Hub-Authority (HITS Algorithm)

- starts with all hub and authority scores equal to 1
- chooses a number of steps K
- performs a sequence of K Authority and Hub updates in this order.

$$a(1) = h(2) + h(3) + h(4)$$



$$h(1) = a(5) + a(6) + a(7)$$



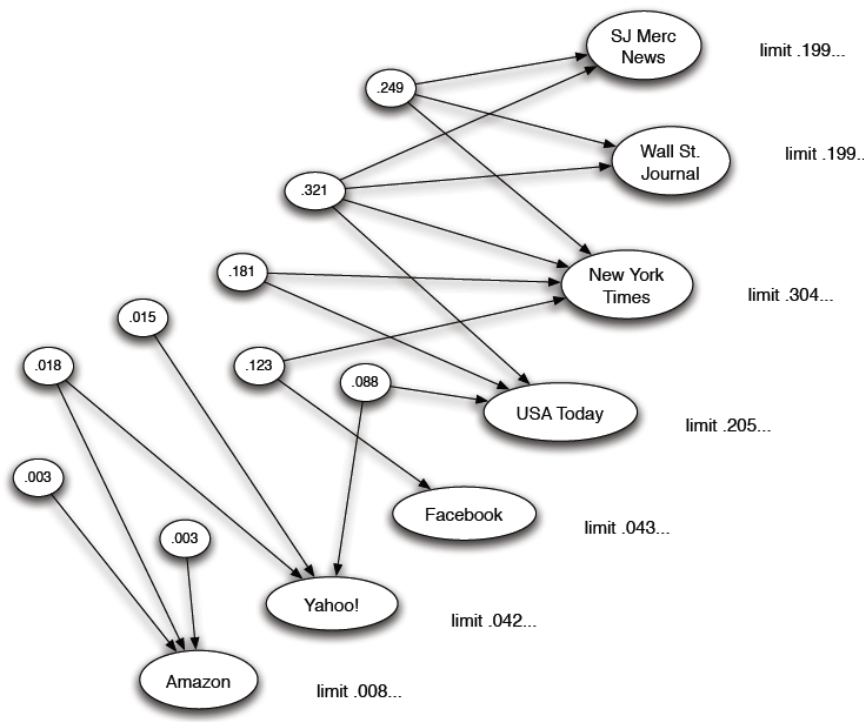
Hub-Authority (HITS Algorithm)

- starts with all hub and authority scores equal to 1
- chooses a number of steps K
- performs a sequence of K Authority and Hub updates in this order.

- Problems
  - Score grows to very large numbers
  - Actually converges?

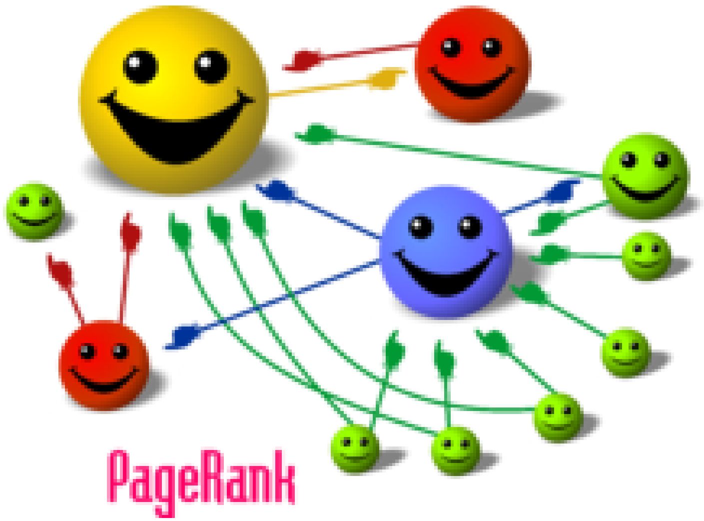
Hub-Authority (HITS Algorithm)

- Problems
  - Score grows to very large numbers
    - normalization
  - Actually converges?
    - Equilibrium
    - Effect of initial values



PageRank

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.
—Facts about Google and Competition



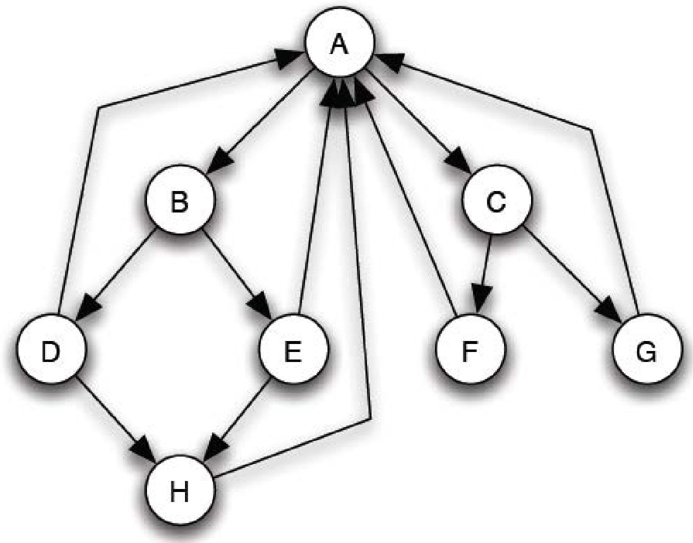
- Keys:
  - Mode of endorsement form the basis of PageRank
  - Starts with simple voting on in-links
  - Pass endorsement across out-links
  - Repeated improvement

PageRank (contd.)

Think as kind of "fluid" that circulates through networks

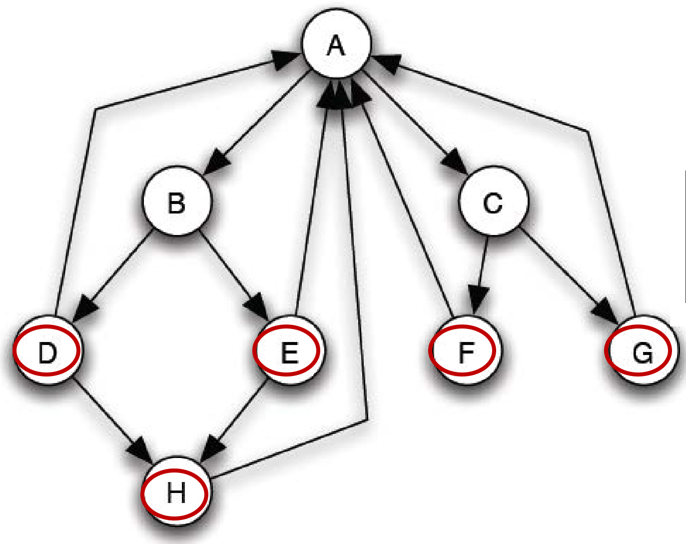
- Computation procedure:
  - Each node with initial pagerank $\frac{1}{n}$
  - A number of steps K
  - K updates of PageRank values
    - Each node/page divides it current PageRank value equally across its out-links
    - Each page updates its new PageRank value to be the sum of what it receives

PageRank (contd.)



What is the PageRank of node A at step 1?

PageRank (contd.)

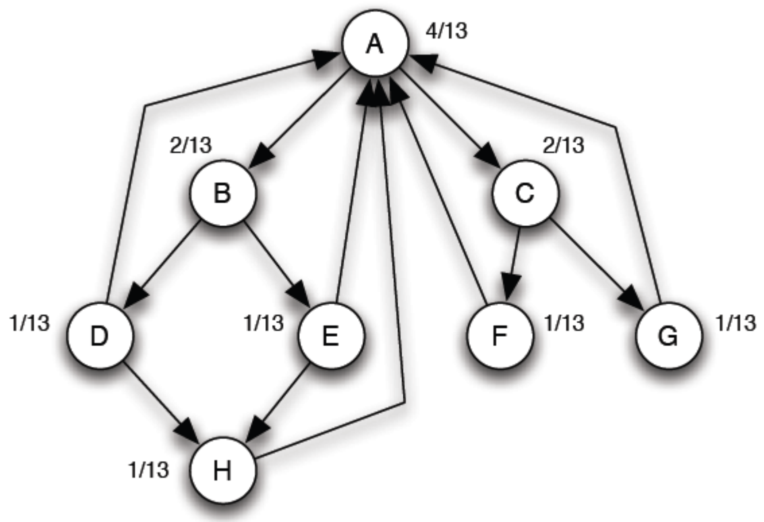


Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

- Computation procedure:
  - Each node with initial pagerank $\frac{1}{8}$
  - Step 1: PR(A) = ½*PR(D) + ½*PR(E) + PR(H) + PR(F) + PR(G)=1/16+1/16+1/8+1/8+1/8=1/2

PageRank (contd.)

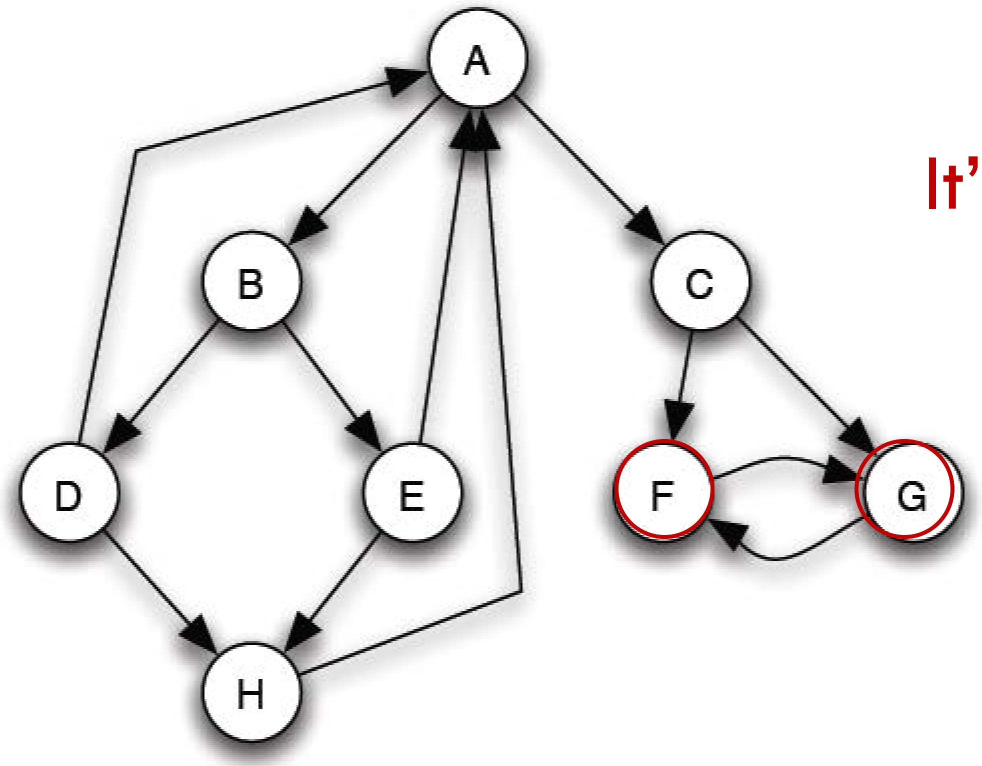
- Convergence/equlibrium?
  - Is there any?
  - How to check?



PageRank (contd.)

- Do you see any problem with the definition?

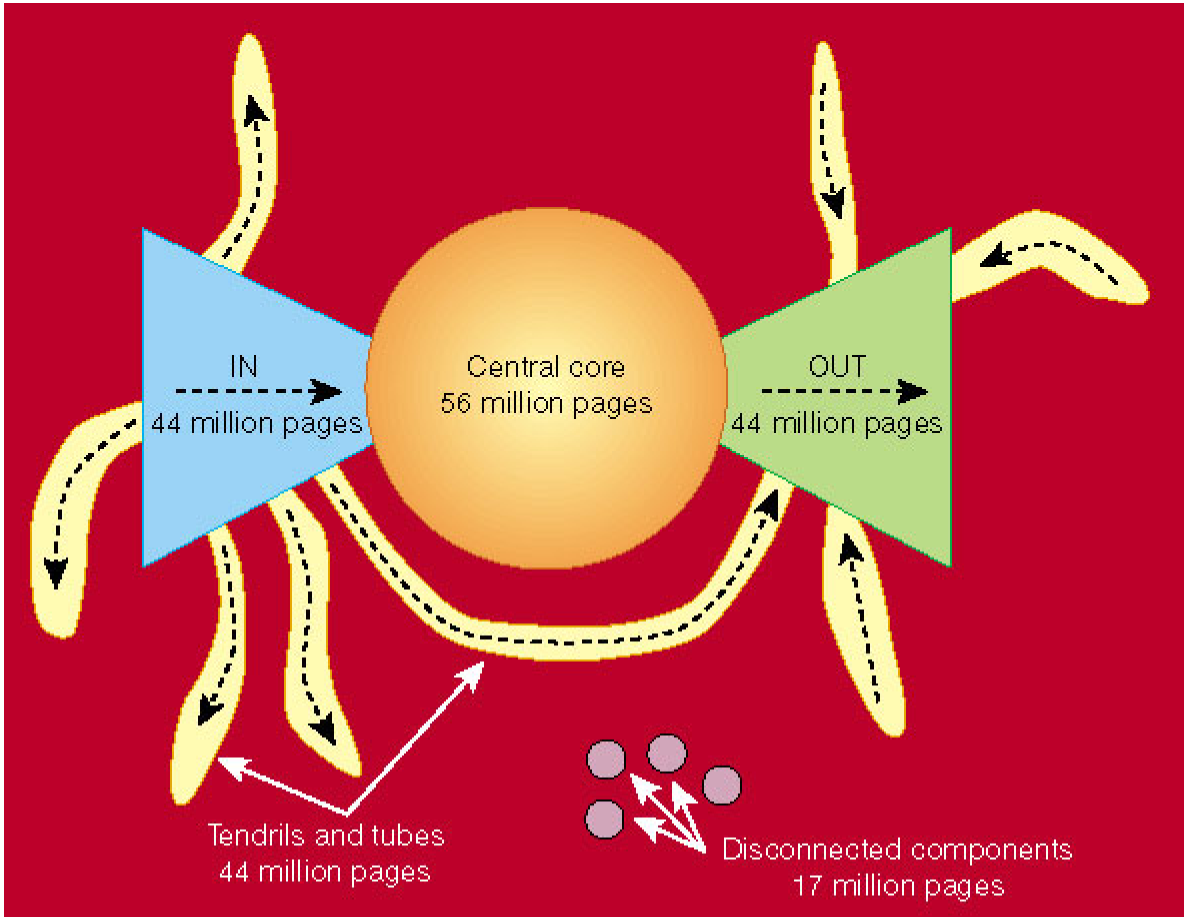
PageRank (contd.)



It’s leaking!

PageRank (contd.)

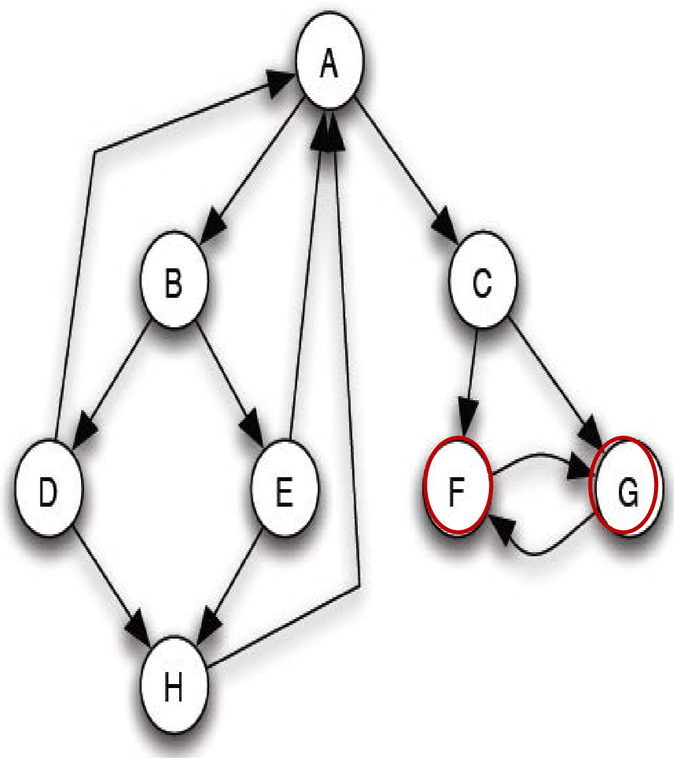
What would happen here? [Broder et al. 2001]



PageRank (contd.)

Solution: scaled PageRank Update rule

- Scaling factor s
- Scale down all PageRank values by a factor of s
- Divide residual 1-s equally over all nodes, (1-s)/n to each.



PageRank (contd.)

Limit of scaled PageRank

- Still converges?
- Depends on scaling factor?
- Sensitivity to addition/deletion of pages?

PageRank (contd.)

Limit of scaled PageRank

- Still converges? YES
- Depends on scaling factor? YES
- Sensitivity to addition/deletion of pages? [Ng et al. 2001]

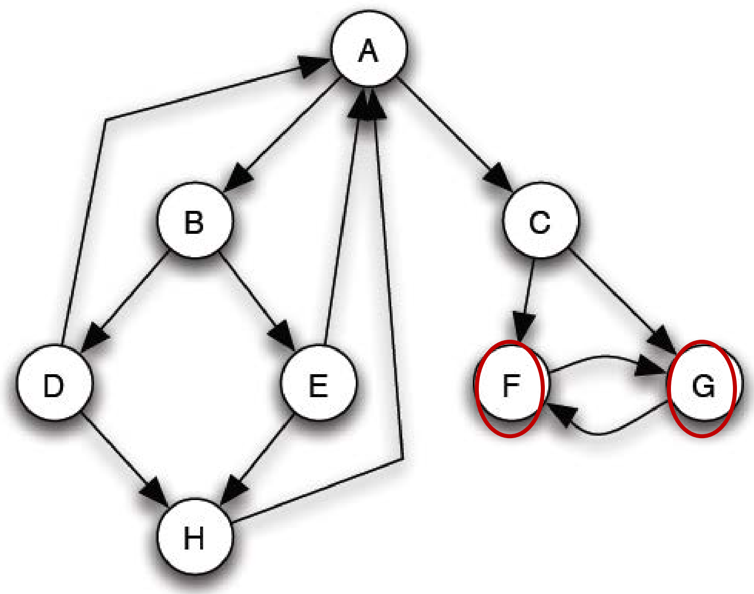
PageRank: alternate definition

Random walk

- Choose a page at random
- Pick each edge with equal probability
- Follow links for a sequence of k steps
  - Pick a random out-links
  - Follow it to where it leads

Claim: The probability of being at a page X after k steps of this random walk is precisely the PageRank of X after k applications of the Basic PageRank Update Rule.

PageRank: alternate definition



Scaled version of Random walk?

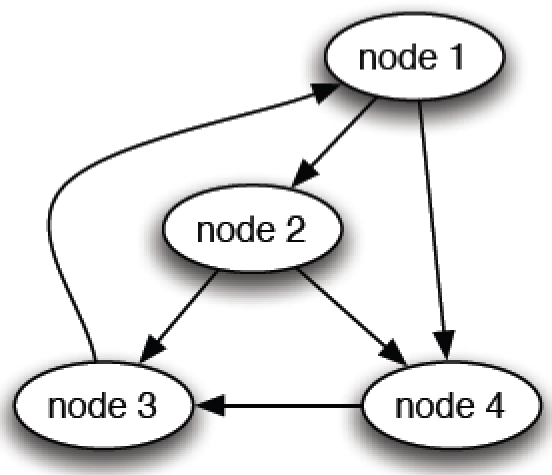
PART 2B

Spectral analysis of Hub-Authority and PageRank

Spectral Analysis of Hub-Authorities

Goal: Hub-authority computation converges to limiting values

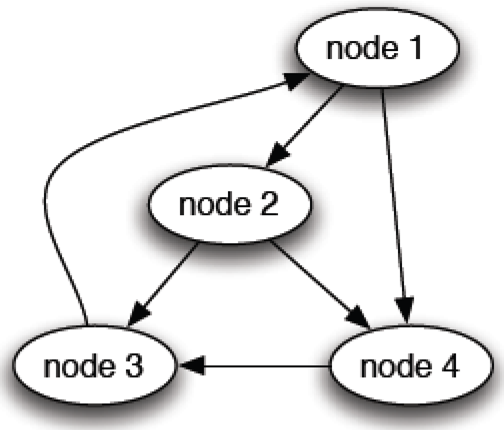
- Adjacency matrix representation of link structure, M_{ij}
- Hub and authority values of nodes are two distinct vectors



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

Spectral Analysis of Hub-Authorities

Example: Updating hub values



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

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

$$h \leftarrow Ma.$$

Spectral Analysis of Hub-Authorities

Example: Updating authority values

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

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

Spectral Analysis of Hub-Authorities

Example: Updating hub and authority values

1

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

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

2

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

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

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 Hub-Authorities

Example: Updating hub and authority values

k

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

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

Multiplying an initial vector by larger and larger power of M^TM and MM^T respectively

Convergence of Hubs

Normalization required for convergence to limit as k goes to infinity

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

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

- Eigenvector $h^{<*>}$
- Eigenvalue c
- The proof reduces to
  - The sequence of vectors $h^{<*>}/c^k$ indeed converges to an eigenvector of MM^T

Convergence of Hubs

Theorem [Ref 268, Kleinberg Book]

Any symmetric matrix A with n rows and n columns has a set of n eigenvectors that are all unit vectors and all mutually orthogonal — that is, they form a basis *for the space* $\mathbf{R}^n$.

- Orthogonal eigenvector: $z_1, z_2, \ldots, z_n$
- Corresponding eigenvalues: $c_1, c_2, \ldots, c_n$
- Assumptions:

$$|c_1| \geq |c_2| \geq \cdots \geq |c_n|$$

$$|c_1| > |c_2|$$

$$x = p_1 z_1 + p_2 z_2 + \cdots + p_n z_n$$

Convergence of Hubs

Proof:

$$
\begin{aligned}
(MM^T)x &= (MM^T)(p_1 z_1 + p_2 z_2 + \cdots + p_n z_n) \\
&= p_1 MM^T z_1 + p_2 MM^T z_2 + \cdots + p_n MM^T z_n \\
&= p_1 c_1 z_1 + p_2 c_2 z_2 + \cdots + p_n c_n z_n,
\end{aligned}
$$

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

In a similar fashion,

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

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

As k goes to infinity, $\frac{h^{\langle k \rangle}}{c_1^k} = q_1 z_1$

Convergence of Hubs

Proof (contd.):

Needs to show that,

- 1. The coefficient q_1 is not zero.
- 2. Limit exists regardless of the initial hub values
  - Any positive initial vector x works; different linear combination.

Proving 1,

$$z_1 \cdot x = z_1 \cdot (p_1 z_1 + \cdots p_n z_n) = p_1(z_1 \cdot z_1) + p_2(z_1 \cdot z_2) + \cdots + p_n(z_1 \cdot z_n) = p_1,$$

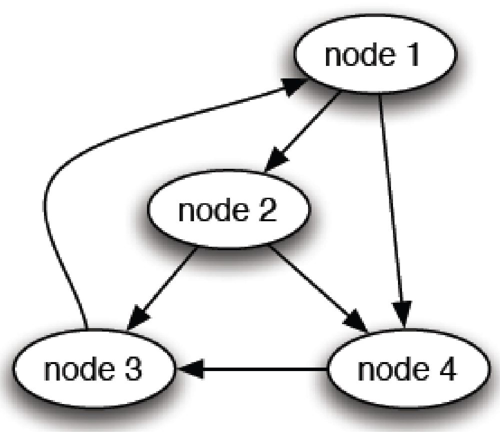
Only requirement is x not being orthogonal to z_1

Can be proved that no positive vector is orthogonal to z_1

Spectral analysis of PageRank

Goal: PageRank computation converges to limiting values

- Adjacency matrix representation of link structure, N_{ij}, portion of i's pagerank that should be passed to j in one update step.
- PageRank vector r



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

Spectral analysis of PageRank

Goal: PageRank computation converges to limiting values

- If l_i outgoing edges: $N_{ij} = 1/\ell_i,$
- If no outgoing edge: $N_{ii} = 1$

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

$$r \leftarrow N^T r$$

Scaled version, $sN_{ij} + (1-s)/n$

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

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

Convergence of PageRank

Proof:

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

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

$\tilde{N}$ that are not symmetric,

We will apply here Perrons Theorem [Ref268, Kleiberg Book]

matrix P in which all entries are positive has the following properties.

(i) P has a real eigenvalue $c > 0$ such that $c > |c'|$ for all other eigenvalues c'.

(ii) There is an eigenvector y with positive real coordinates corresponding to the largest eigenvalue c, and y is unique up to multiplication by a constant.

(iii) If the largest eigenvalue c is equal to 1, then for any starting vector $x \neq 0$ with non-negative coordinates, the sequence of vectors $P^k x$ converges to a vector in the direction of y as k goes to infinity.

PageRank as a probability of random walk

$$b_i \leftarrow N_{1i}b_1 + N_{2i}b_2 + \cdots + N_{ni}b_n.$$

$$b \leftarrow N^T b.$$

Scaled version,

$$b_i \leftarrow \tilde{N}_{1i}b_1 + \tilde{N}_{2i}b_2 + \cdots + \tilde{N}_{ni}b_n$$

$$b \leftarrow \tilde{N}^T b$$

Claim: The probability of being at a page X after k steps of the scaled random walk is precisely the PageRank of X after k applications of the Scaled PageRank Update Rule.

Questions?