Link Analysis and Web Search (PageRank)

Endorsement in HITS

- Links denote (collective) endorsement
- Multiple roles in the network:
  - Hubs: pages that play a powerful endorsement role without themselves being heavily endorsed
  - Authorities: pages being heavily endorsed

- Why separate Hubs from Authorities?
  - Competing firms will not link to each other
  - Can't be viewed as directly endorsing each other

Endorsement in PageRank

- Endorsement passes directly from one prominent page to another
- A page is important if it is cited by other important pages
  - dominant mode of endorsement in academic or governmental pages, among bloggers, among personal pages, or in scientific literature (pdf's)

Summary of PageRank

- Start with simple voting based on in-links
- Refine it with repeated improvement
  - nodes repeatedly pass endorsements across their out-going links
  - the weight of a node's endorsement is based on the current estimate of its PageRank
- Nodes that currently viewed as more important make stronger endorsements

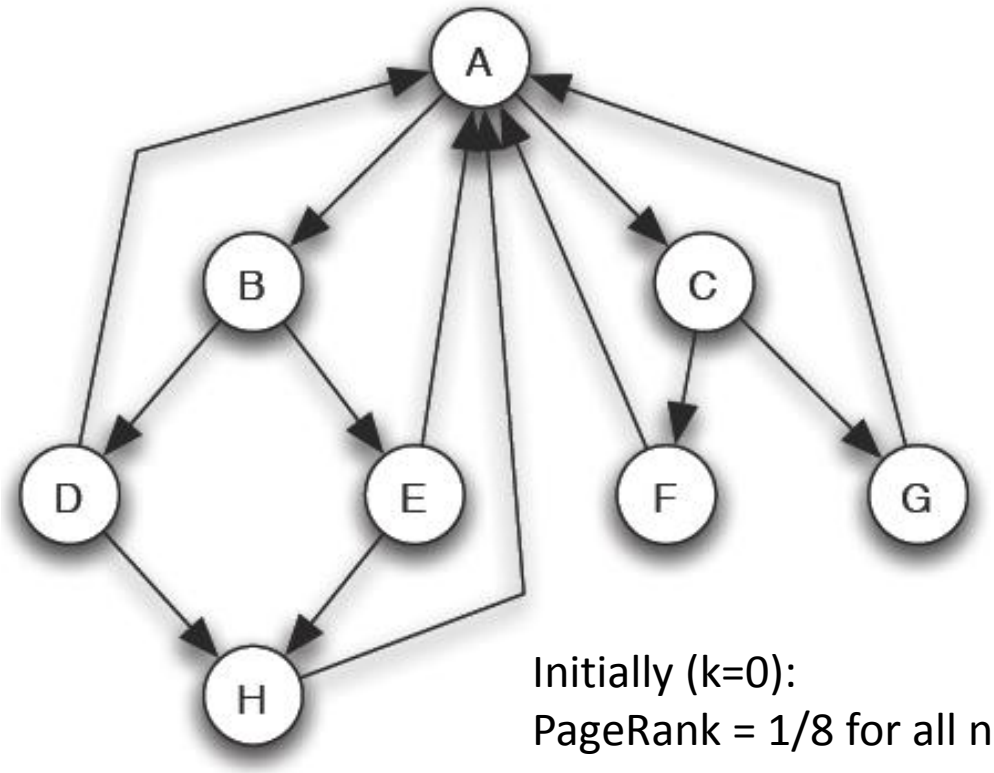
Basic definition of PageRank

- In a network with *n* nodes, we assign all nodes the same initial PageRank, set to be 1/*n*
- We choose a number of steps *k*
- We then perform a sequence of *k* updates to the PageRank values, using the following rule for each update
- Basic PageRank Update Rule:
  - Each page divides its current PageRank equally across its out-going links, and passes these equal shares to the pages it points to
  - (If a page has no out-going links, it passes all its current PageRank to itself)
  - Each page updates its new PageRank to be the sum of the shares it receives

Intuition in PageRank

- PageRank is a kind of “fluid”:
  - It circulates through the network
  - Is passing from node to node across edges
  - Is pooling at the nodes that are the most important
- Total PageRank in the network remains constant
  - Why? Each page takes its PageRank, divides it up, and passes it along links
  - PageRank is never created nor destroyed, just moved around from one node to another
- No need to normalize PageRank of nodes to prevent them from growing
  - Unlike HITS

Example



Initially (k=0):
PageRank = 1/8 for all nodes

A gets a PageRank of 1/2 after the first update because it gets all of F's, G's, and H's PageRank, and half each of D's and E's. On the other hand, B and C each get half of A's PageRank, so they only get 1/16 each in the first step. But once A acquires a lot of PageRank, B and C benefit in the next step

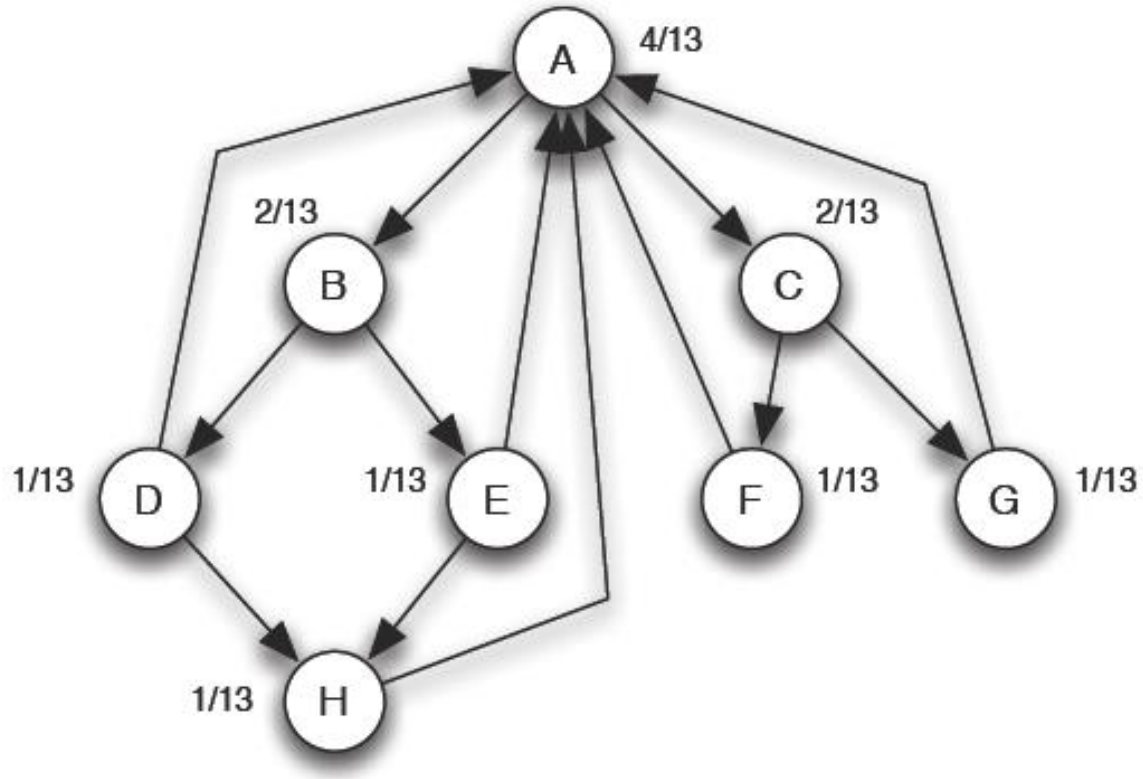
Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

Equilibrium Values of PageRank

- PageRank values of all nodes converge to limiting values as the number of update steps *k* goes to infinity (except in certain degenerate special cases)
- If the network is strongly connected, then there is a unique set of equilibrium values

- Interpretation of limit: by applying one step of the Basic PageRank Update Rule, the values at every node remain the same (i.e., regenerate themselves exactly when they are updated)

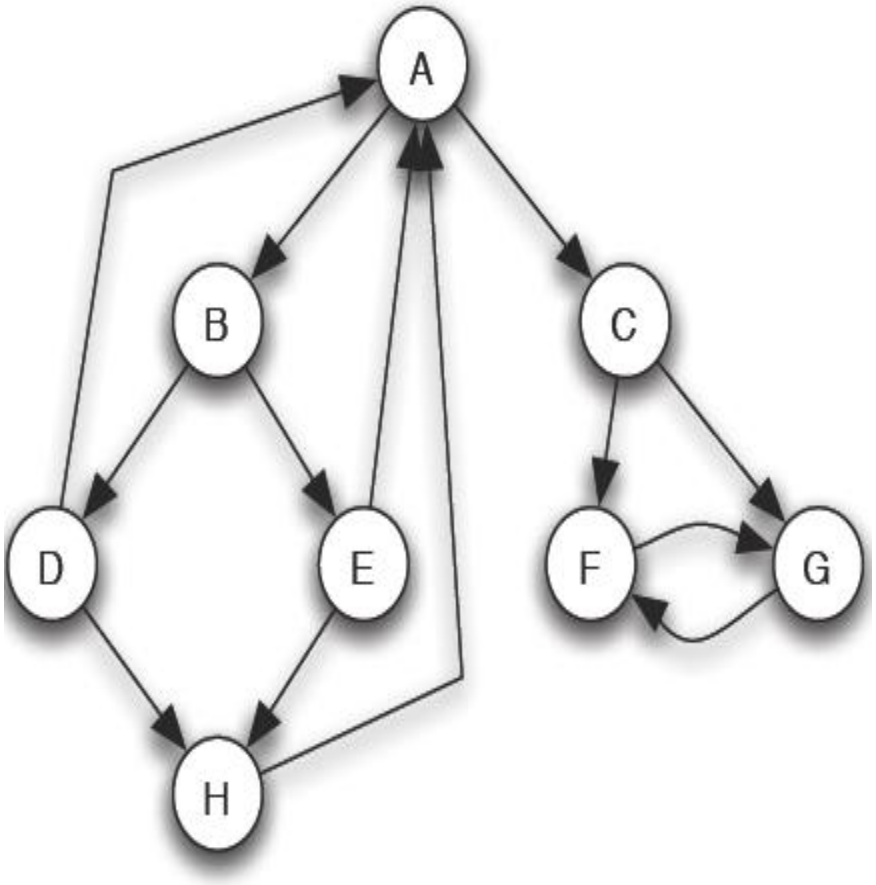
Example



Equilibrium PageRank values

Problem with Basic Definition of PageRank

- In many networks, the "wrong" nodes can end up with all the PageRank
- Ex (figure):
  - F and G point to each other rather to A
  - PageRank that flows from C to F and G can never circulate back: for large *k* we have 1/2 for each of F and G, and 0 for all other
  - "slow leak": mall sets of nodes that can be reached from the rest of the graph, but have no paths back



Solution to the problem

- Remember 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

Scaling the definition of PageRank

- Pick a scaling factor *s* between 0 and 1
- Replace the Basic PageRank Update Rule with the following
- Scaled PageRank Update Rule:
  - First apply the Basic PageRank Update Rule
  - Scale down all PageRank values by a factor of *s*
    - total PageRank in the network has shrunk from 1 to *s*
  - Divide the residual 1 − *s* units of PageRank equally over all nodes, giving (1 − *s*)/*n* to each
- Why it works?
  - "water cycle" that evaporates 1 − *s* units of PageRank in each step and rains it down uniformly across all nodes

The Limit of the Scaled PageRank Update Rule

- Repeated application of the Scaled PageRank Update Rule converges to a set of limiting PageRank values as the number of updates *k* goes to infinity
- Unique equilibrium
  - But values depend on our choice of the scaling factor *s*
- The version of PageRank used in practice, with a scaling factor *s* between 0.8 and 0.9

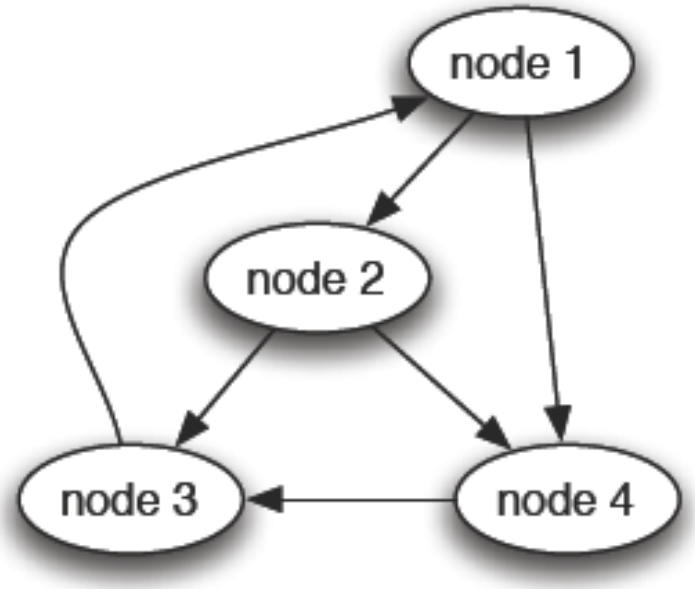
Spectral Analysis of PageRank

- How PageRank can be analyzed using matrix-vector multiplication and eigenvectors?
- Start with Basic PageRank Update Rule and then move on to the scaled version
- Approach similar to HITS
  - no need for normalizing

Spectral Analysis of PageRank

- The "flow" of PageRank represented using a matrix *N*
- Define N_{ij} *to be the* share of *i*'s PageRank that *j* should get in one update step
  - $N_{ij} = 0$ if *i* doesn't link to *j*
  - Else N_{ij} is the reciprocal of the number of nodes that *i* points to
    - If *i* has no outgoing links, then we define $N_{ij} = 1$

Example



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

Spectral Analysis of PageRank

- Represent PageRanks of all nodes using a vector *r*
- *Define* r_i to be the PageRank of node *i*
- Write the Basic PageRank Update Rule as

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

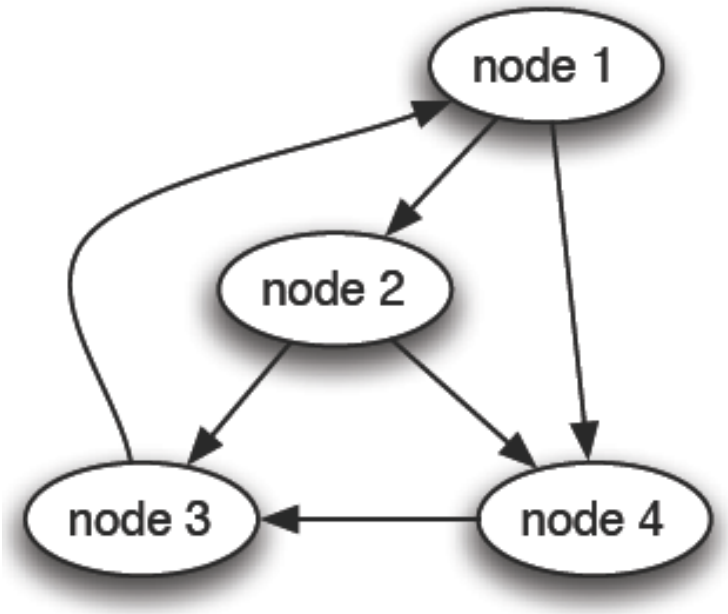
- This corresponds to multiplication by the transpose of the matrix

$$r \leftarrow N^T r$$

Spectral Analysis of PageRank

- Scaled PageRank Update Rule represented in the same way, but with a different matrix $\tilde{N}$ to represent the different flow of PageRank
- Define $\tilde{N}_{ij} = sN_{ij} + (1-s)/n$

Example



$$\begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$

$s = 0.8$

Spectral Analysis of PageRank

- The scaled update rule can be written as

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

- Or equivalently

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

Repeated Improvement Using the Scaled PageRank Update Rule

- Starting from an initial PageRank vector $r^{<0>}$ we produce a sequence of vectors $r^{<1>}$, $r^{<2>}$, . . .
- We see that

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

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

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

- This happens when

$$r^{\langle * \rangle} \text{ an eigenvector of } \tilde{N}^T$$

Convergence of the Scaled PageRank Update Rule

- In HITS, the matrices involved (MM^T and M^TM) were symmetric, and so they had eigenvalues that were real numbers
- In general, $\tilde{N}$ is not symmetric, but $\tilde{N}_{ij} > 0$
- *Perron's Theorem:* for matrices *P* such with all entries positive
  - *P* has a real eigenvalue $c > 0$ such that $c > |c'|$ for all other eigenvalues c'
  - 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
  - If the largest eigenvalue *c* is equal to 1, then for any starting vector $x \neq 0$ with nonnegative coordinates, the sequence of vectors $P^k x$ converges to a vector in the direction of *y* as *k* goes to infinity
- This vector *y* corresponds to the limiting PageRank values

Random walks: An equivalent definition of PageRank

- Random walk on the network
  - start by choosing a page at random (each page with equal probability)
  - follow links for a sequence of k steps:
  - in each step, pick a random out-going link from the current page
  - (if the current page has no out-going links, stay where you are) Such an exploration of

Random walks: An equivalent definition of PageRank

- The probability of being at a page X after k steps of a random walk is precisely the PageRank of X after k applications of the Basic PageRank Update Rule
- The PageRank of a page X is the limiting probability that a random walk across hyperlinks will end up at X, as we run the walk for larger and larger numbers of steps

Formulation of PageRank Using Random Walks

- $b_1, b_2, \ldots, b_n$ denote the probabilities of the walk being at nodes 1, 2, . . . , *n* respectively in a given step
- Write the update to the probability b_i:

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

- This is equal to:

$$b \leftarrow N^T b$$

PageRank values and random-walk probabilities start the same (initially 1/*n*) and are updated with the same rule

A Scaled Version of the Random Walk

- With probability *s,* the walk follows a random edge as before; and with probability 1 – *s* it jumps to a node chosen uniformly at random
- Write the update to the probability b_i*:*

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

- This is equal to:

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

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

Applying Link Analysis in Web Search

- The link analysis ideas played an integral role in the ranking functions of Google, Yahoo!, Microsoft's search engine Bing, and Ask
- Link analysis ideas have been extended and generalized considerably
  - combine text and links for ranking is through the analysis of *anchor text* (weight more links with relevant anchor text)
  - Click-through statistics
- Search engine companies themselves are extremely secretive about what goes into their ranking functions
  - also to protect themselves from SEO