CS 103: Lecture 12 Link Analysis for Web Search

Dan Sheldon

November 18, 2015

Announcements

- HW 3 back today
- HW 4 due today
- Midterm Tuesday
- Guest lecture next Thursday in Cleveland L2

Midterm: Topics

- Graph Theory
- Strong and Weak Ties
- Signed networks and structural balance
- Game theory
- Braess's paradox / traffic in networks
- Auctions
- Matching markets
- Network exchange

Be able to do problems like those on your homework and answer short conceptual questions about these topics

Midterm: What You Don't Need to Know

Web Search

Web search is hard! Some history:

- Information retrieval ca. 1960s
- Keyword search of curated collections (libraries, patents, etc.)
- "Inverted index"
- Challenges
  - **synonymy**: two words, one meaning
    - green onions vs. scallions
  - **polysemy**: one word, two meanings
    - Yosemite (Mac OS) vs. Yosemite (National Park)
- Try this: "window installation" vs. "install windows"

Web Search

The Web made a hard problem harder

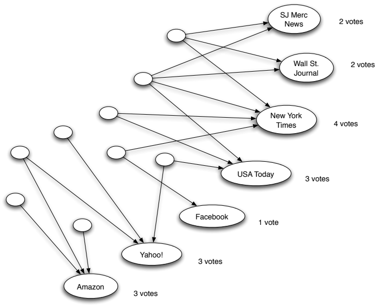
- Huge diversity of documents
- E.g., millions of documents relevant to "Holyoke"
  - MHC home, US News and World Report, Mount Holyoke State Park, City of Holyoke, Pages about alums, etc.

How to find *best* or *most authoritative documents*?

- Link-analysis (late 1990s)
  - Hubs and Authorities (Kleinberg)
  - PageRank (Google)

Hubs and Authorities

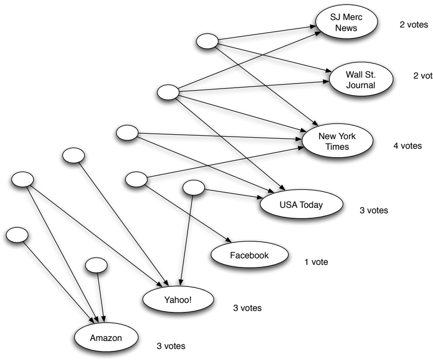
E.g., query "newspaper"



- first use text-based retrieval to get a set of relevant documents
- then use links among them to determine which are authoritative

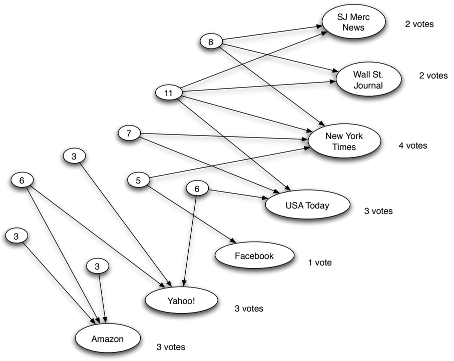
Hubs and Authorities

Step 1: an inlink is a vote for a page



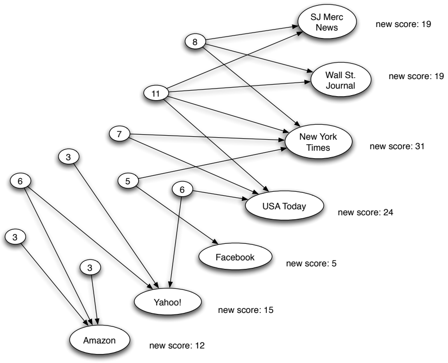
Hubs

Step 2: pages that link to more authoritative sites are better information brokers ("hub score")



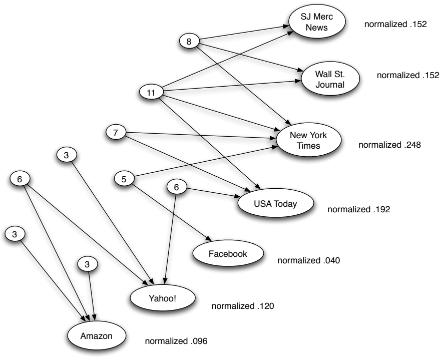
Authorities

Step 3: update authority scores as sum of hub scores from linking pages



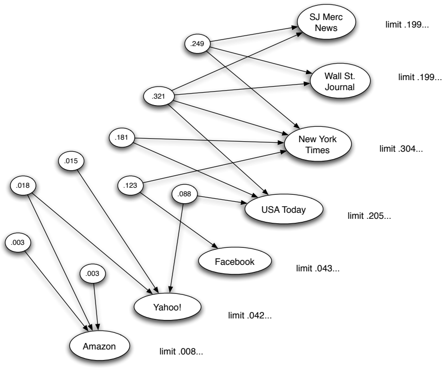
Normalization

Problem: scores are getting very big. Let's normalize them to sum to one.



Wash, Rinse, Repeat

If we repeat forever, this is what we get:



Hubs and Authorities Algorithm

Assign initial hub and authority scores. For each page p:

- Set hub(p) = 1
- Set auth(p) = 1

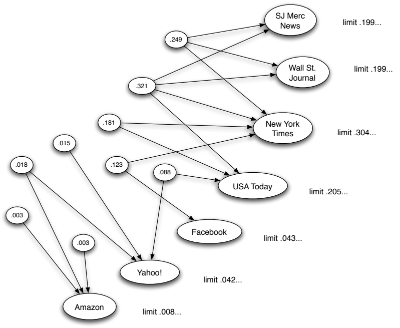
Repeat for k steps

- *Authority update*: for each page p, update auth(p) to be the sum of the hub scores of all pages that point to it
- *Hub update*: for each page p, update hub(p) to be the sum of the authority scores of all pages that it points to

Normalize authority scores to sum to one

Hubs and Authorities Algorithm

It can be shown using linear algebra (eigenvectors/eigenvalues) that this process converges to a unique answer as k goes to infinity:



PageRank

Another link analysis algorithm. Similar principles to Hubs and Authorities, but several key differences:

- Runs on entire web
- Only one type of page
- Each page has a single vote that is divided equally among pages it points to

Basic PageRank

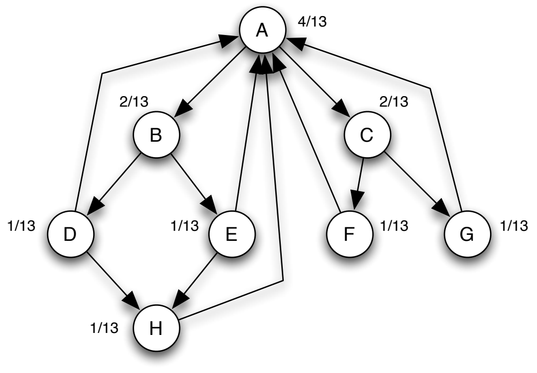
Intution: "fluid" or "currency" passing from node to node in a directed graph

Example on board

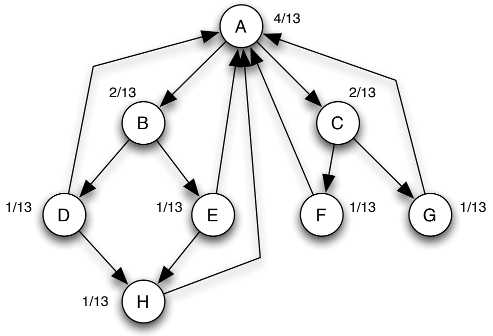
- Assign each node initial PageRank of $1/n$, where $n = \#$ nodes
- **Basic PageRank Update** (repeat k times)
  - Each page divides current PageRank value equally across outgoing links and passes these shares to its neighbors.
  - If a page has no outgoing links, it keeps its current PageRank
  - New PageRank = sum of the shares it receives

Basic PageRank

Exercise: run one PageRank update from this configuration



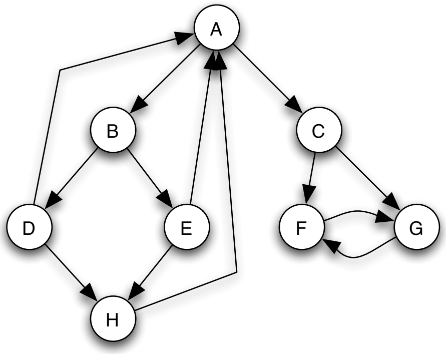
Basic PageRank



If you run PageRank long enough, it will converge to *equilibrium values*, unless. . .

Problem for Basic PageRank

What happens if we keep applying PageRank updates in this graph?



Scaled PageRank

Example on board

- First apply Basic PageRank update
- Then shrink all values by a factor of s
- Now the total "currency" in the network is $1 - s$
- Distribute the remaining s equally among the nodes

Discussion

- Use of PageRank over the years
- Other applications of PageRank
- Manipulation