# *WWW graph?*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- Nodes are not computers but web-pages
- Edges are defined by hyperlinks; they are *directed* from the pointing page to the pointed page

- Precursors:
  - Semantic networks / Mindmaps
  - Associative memory / Memex [V. Bush]

# *Link Analysis*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- Use network structure to answer important problems
  - Search for good web pages
  - Impact of scientific articles via citation analysis
  - Use of court judgments as precedents

# *Web Search*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- Information Retrieval problem
  - Search for ‘jaguar’
  - Search for ‘brinjal’
  - Search for ‘Ebola’ in October 2014
- How can the network structure of the hyperlinks help in this task?

# *Searching the Web: Why is it hard?*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- The web is huge and growing: Google's index of web pages crossed the 1 trillion page mark in July 2008!
- Dynamic: 40% of all webpages change within a week and 23% of .com pages change daily.
- Self-organized: any one can post a webpage and link away at will.
- Data is heterogeneous, and exists in multiple formats, languages and alphabets.

# *Searching the Web: Building blocks*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- *Crawler module:* "Spiders" that scour the web gathering new information and webpages.
- *Indexing module:* Takes webpages and extracts their vital information to create a compressed version of the page. Context is stored in an inverted file structure.
- *Query module:* Converts natural language to numbers and consults content index and its inverted file to find pages that use the query terms
- *Ranking module:* Takes relevant pages and ranks them to produce ordered list of webpages.

# *Page Rank*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- Think of a hyperlink as a recommendation or a vote: a hyperlink from my page to yours is an endorsement of your webpage.
- A page with more recommendations (i.e., inbound links) is more important. As of Nov 2011:
  - www.cmu.edu had 188,000 inbound links (Source: Yahoo! Site Explorer)

# *Page Rank*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- Think of a hyperlink as a recommendation or a vote: a hyperlink from my page to yours is an endorsement of your webpage.
- A page with more recommendations (i.e., inbound links) is more important. As of Nov 2011:
  - Walmart.com had 1.9M inbound links
  - Amazon.com had 220M inbound links
- But the status of the recommender matters as well.

# *Page Rank*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

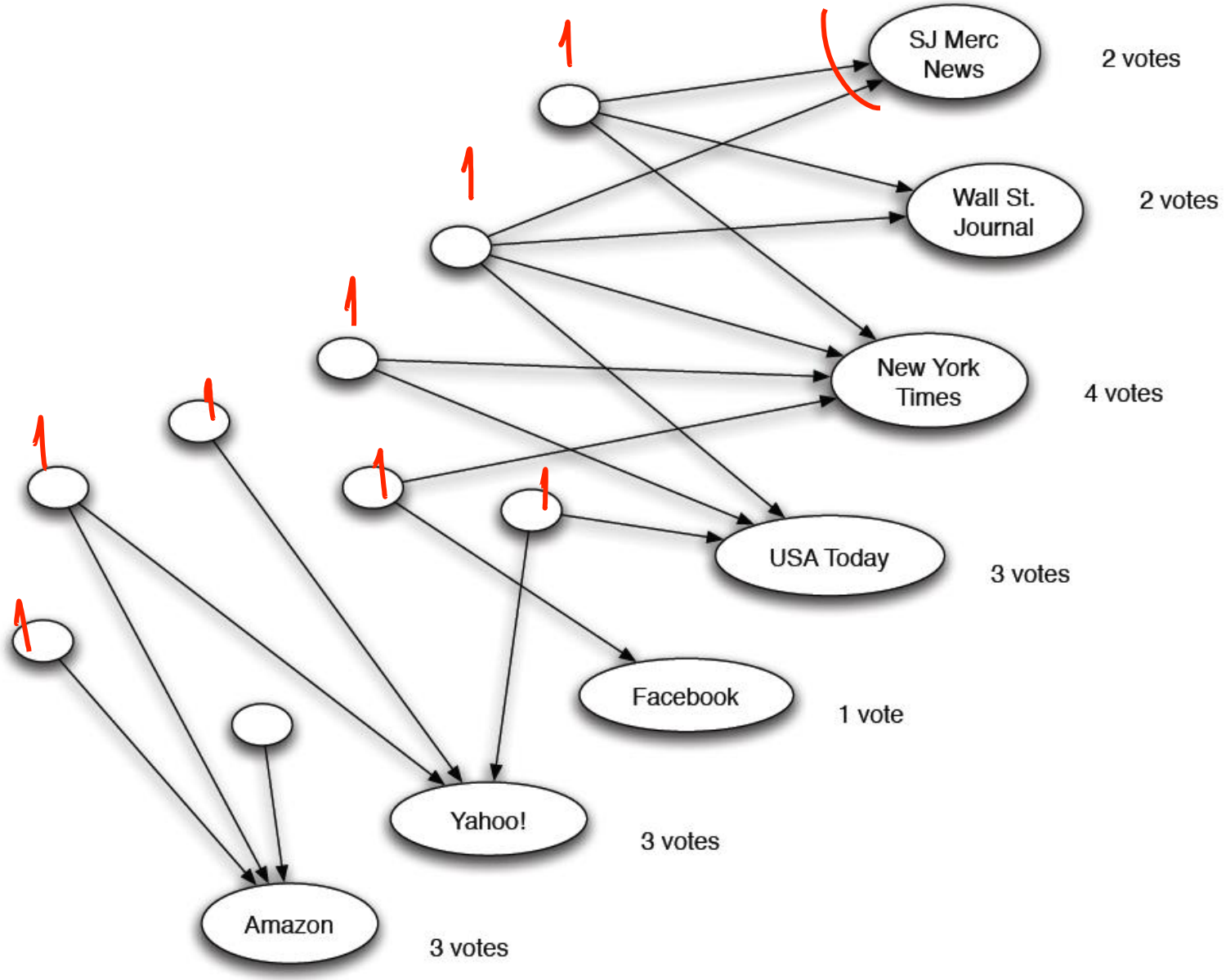
- An endorsement from Warren Buffet probably does more to strengthen a job application than 20 endorsements from 20 unknown teachers and colleagues.
- However, if the interviewer knows that Warren Buffet is very generous with his praise and has written 20,000 recommendations, then the endorsement drops in weight.

# *Initial Abstraction*

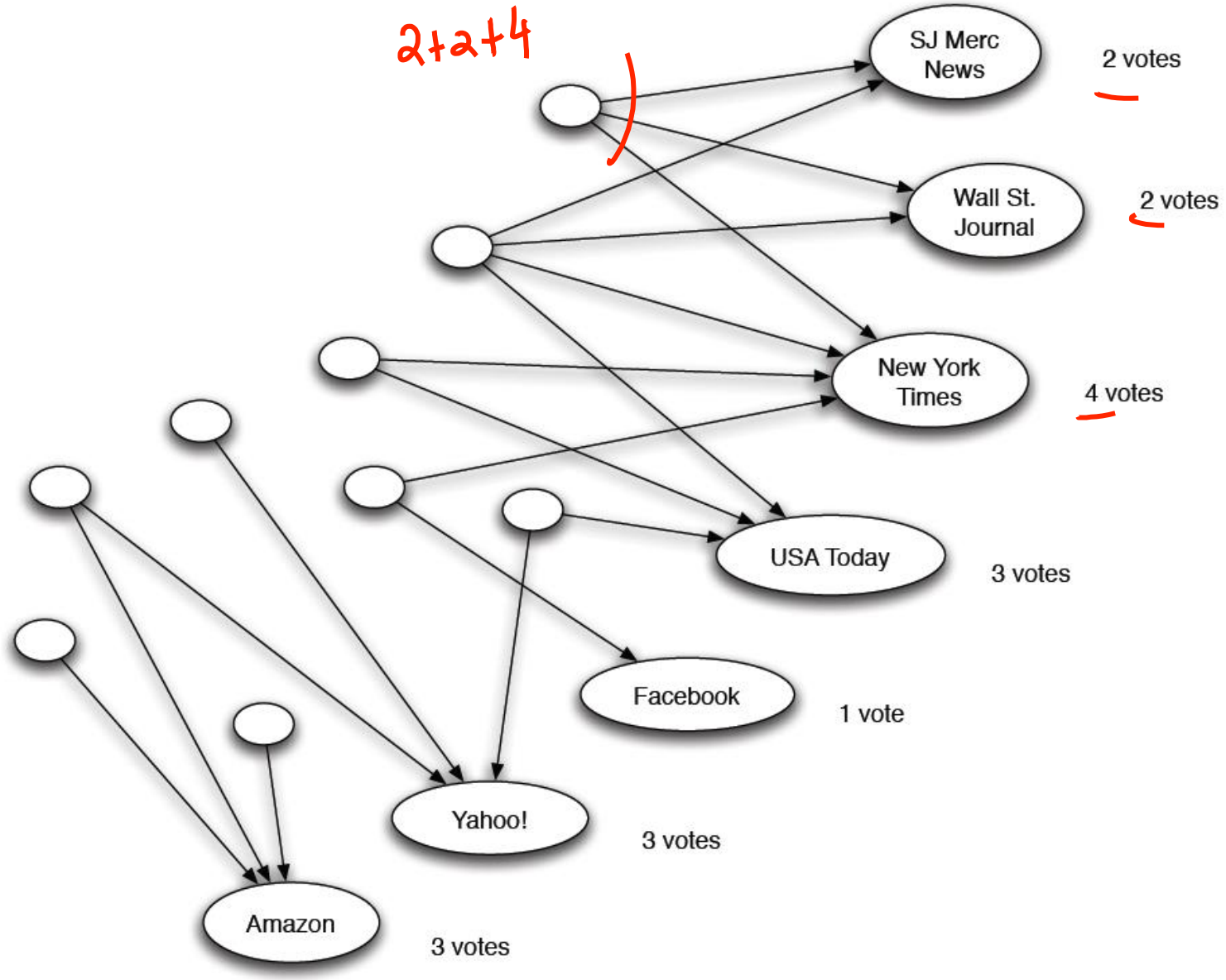
Carnegie Mellon Tepper SCHOOL OF BUSINESS

- Collect all relevant pages for a search query (e.g. that contain the search term or its synonyms)
- Running example for the query "newspapers"

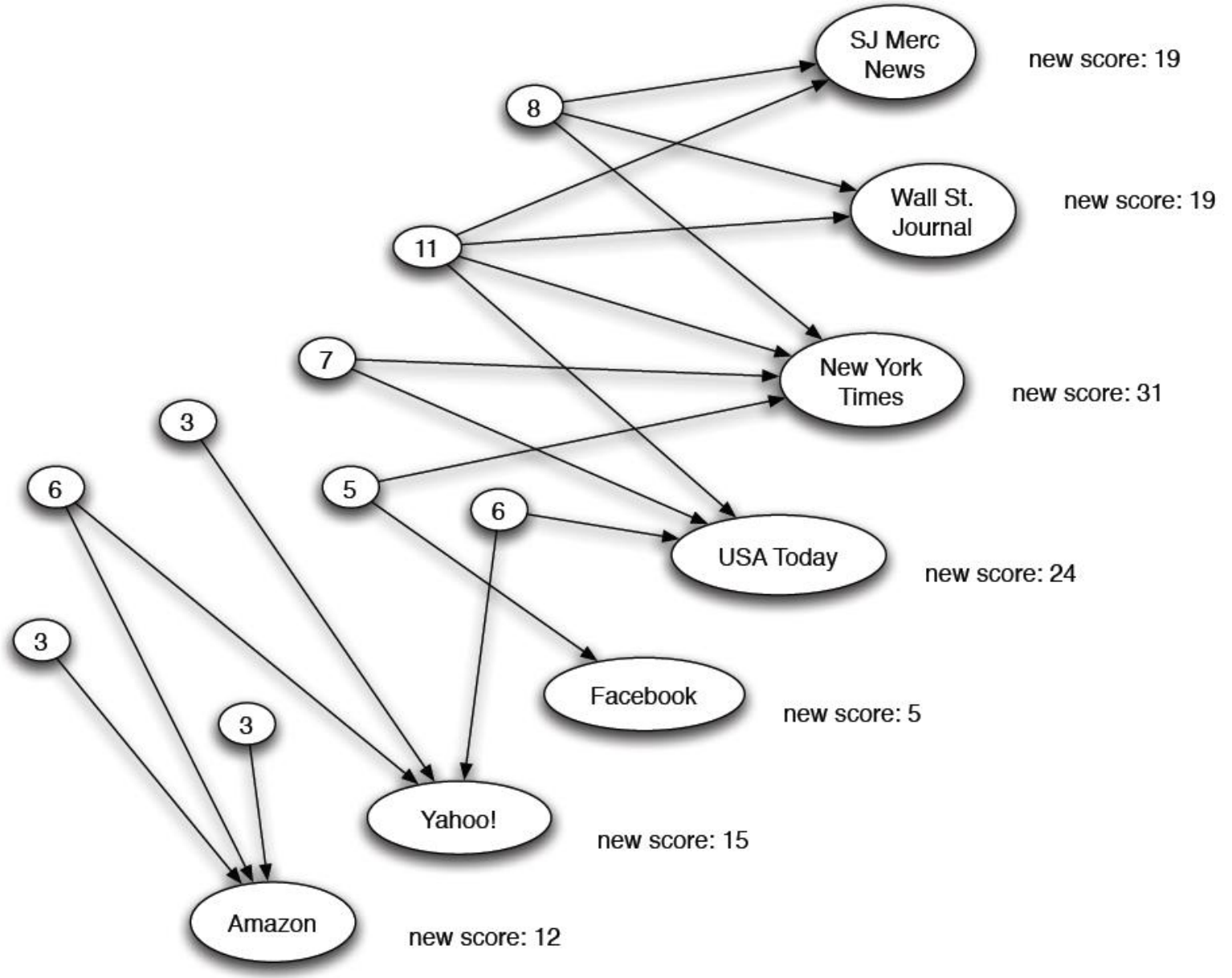
# Voting authorities by in-links



# List-finding hubs by out-links



# *New authority scores*



# *Repeated improvement*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

Initialize all scores to 1

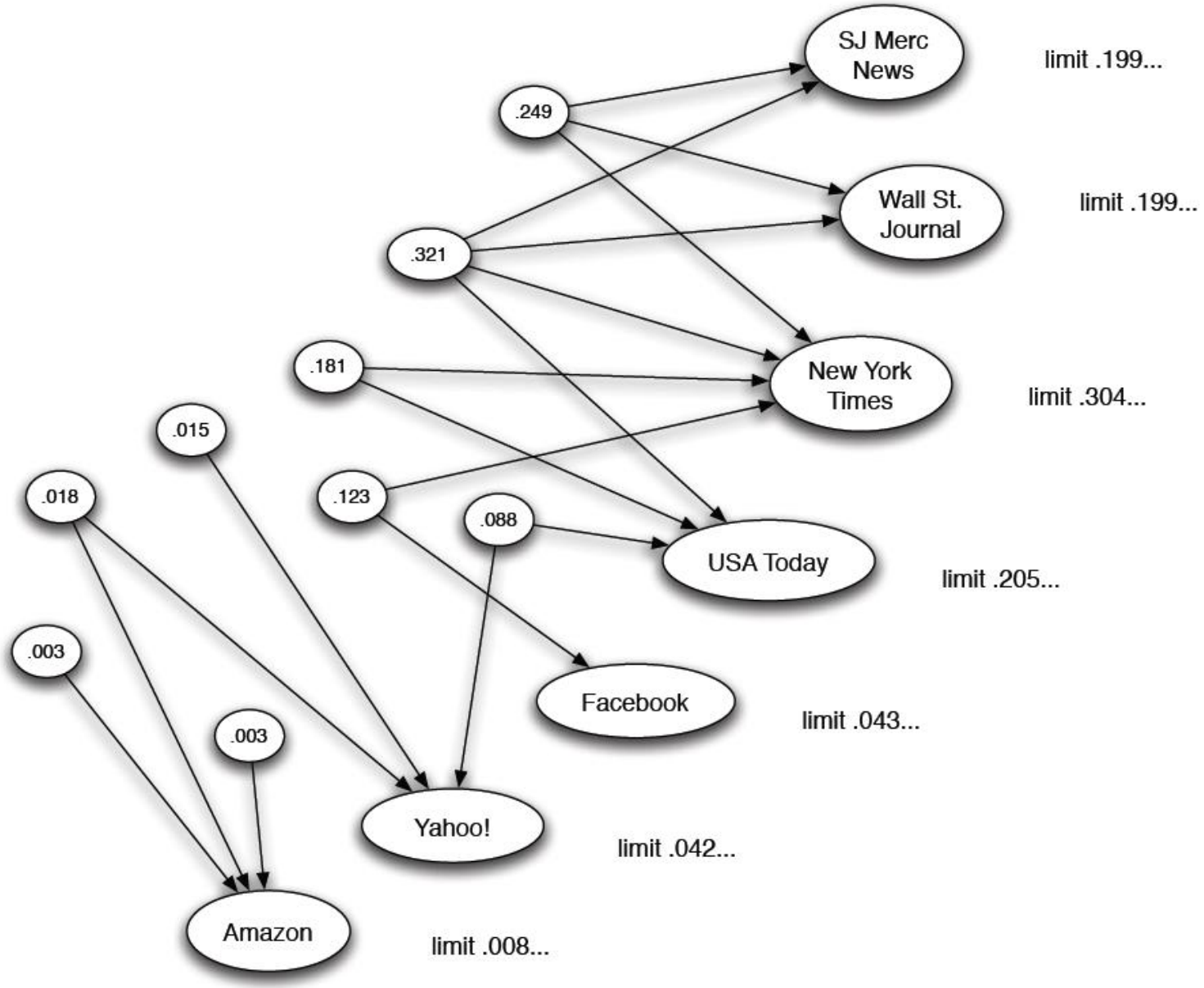
- Repeat for sufficiently many steps k
  - Authority update rule: Update auth(p) to be the sum of the hub scores of all pages pointing to it
  - Hub update rule: Update hub(p) to be sum of the authority scores of all pages it points to
- Normalize by dividing by sum of scores

# *Scores converge*

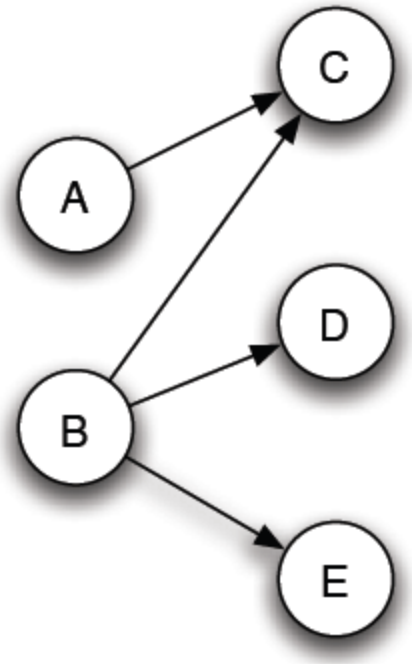
Carnegie Mellon Tepper SCHOOL OF BUSINESS

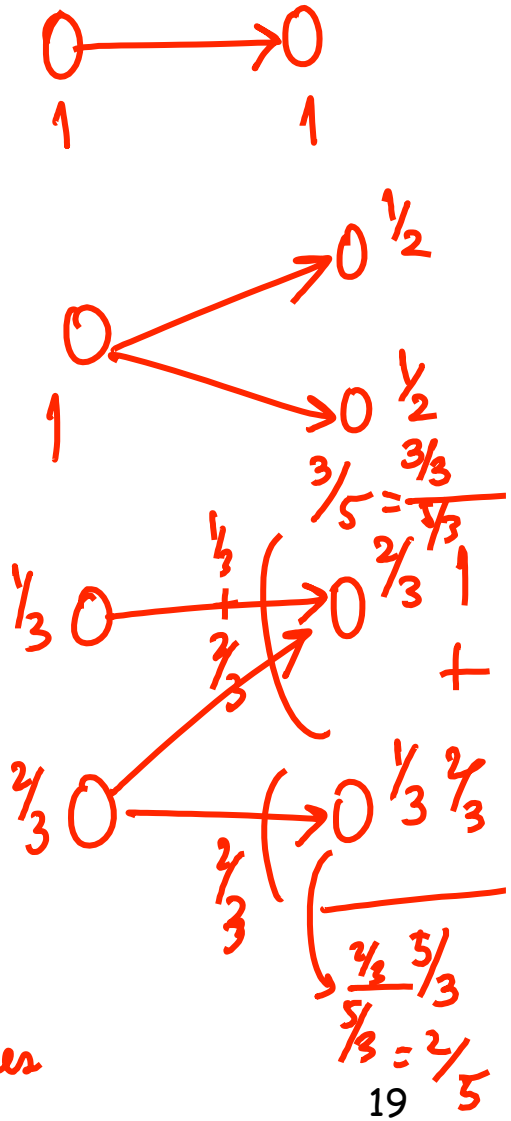
- For larger and larger k
  - Normalized values of hub(p) and auth(p) converge to stable values
  - The values they converge to do not depend on the initial scores (except for some degenerate cases)
  - At converged scores
    - Hub score is proportional to authority of pages it is pointing to
    - Auth score is proportional to hub scores pointing to it

# Scores converge



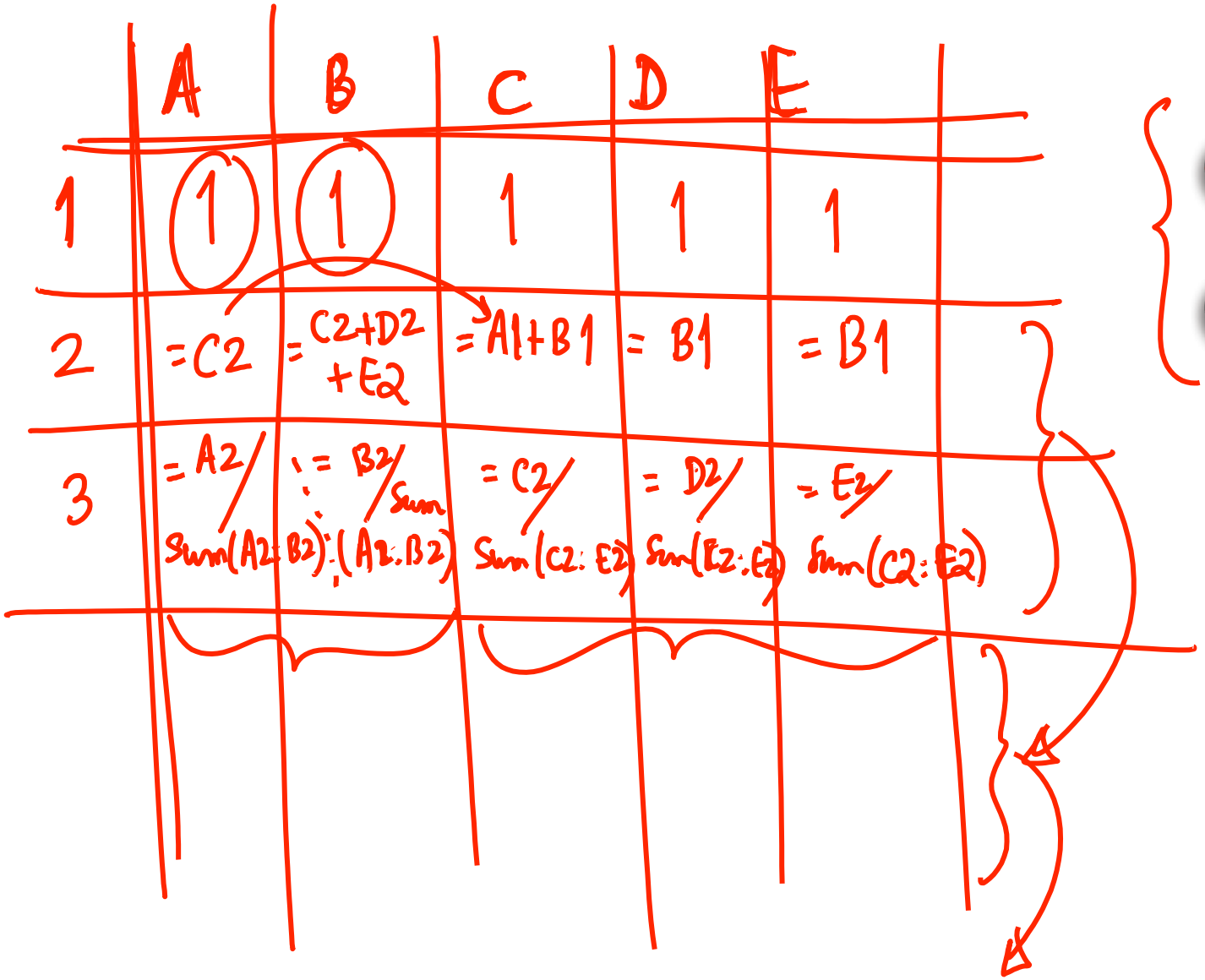
# Simple example

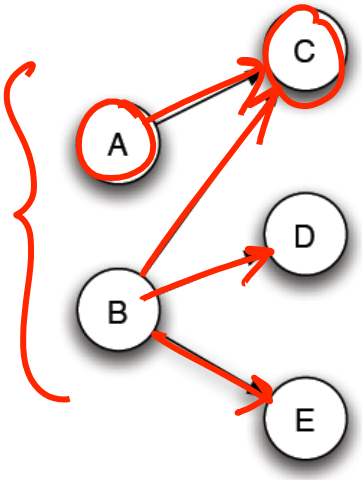


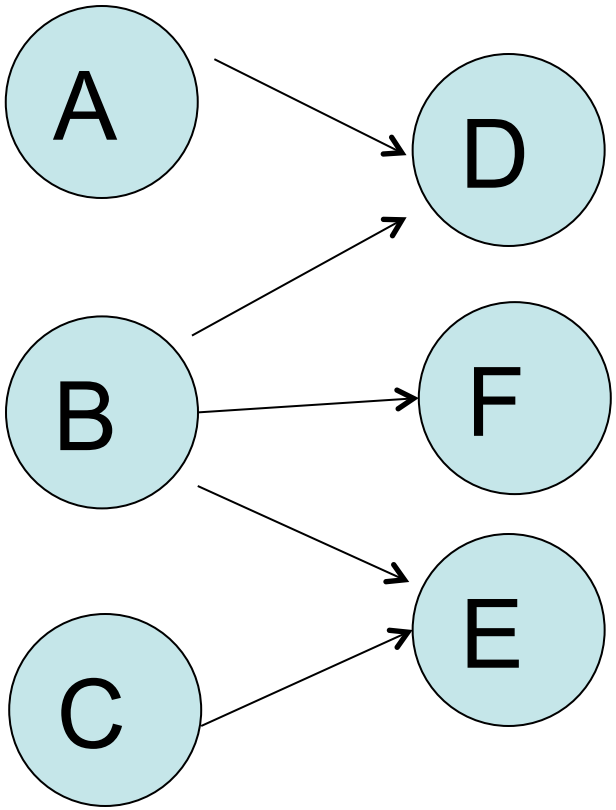

Normalized scores

# Spreadsheet setup?

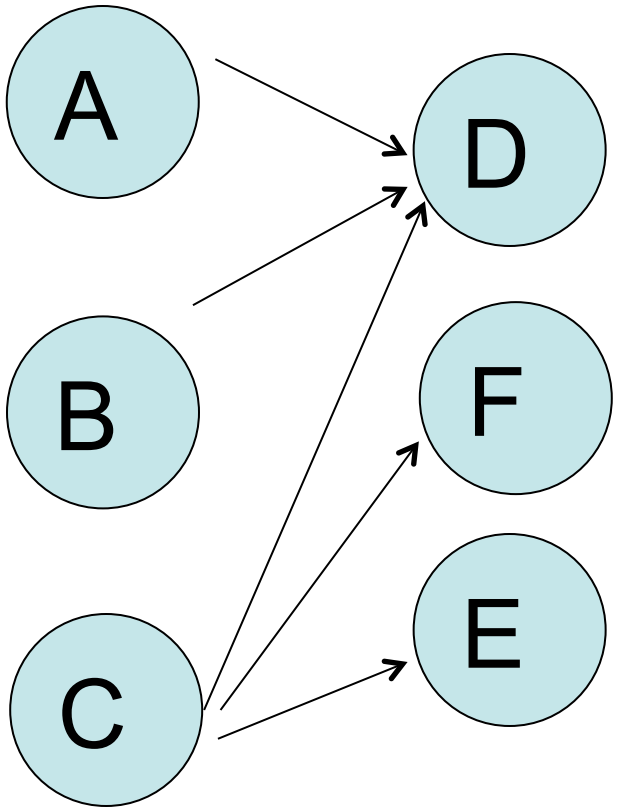


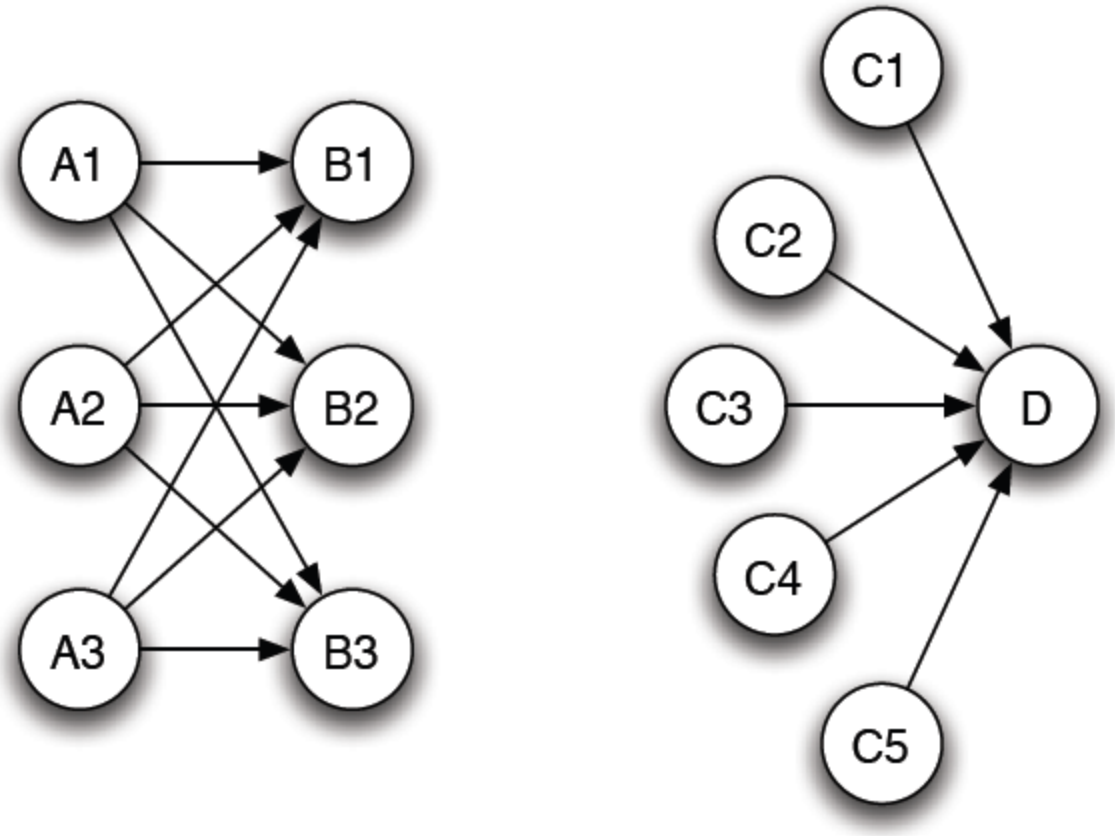
# *Another simple example in video*



# *Try this example*



# *Which will have higher score: B2 or D?*



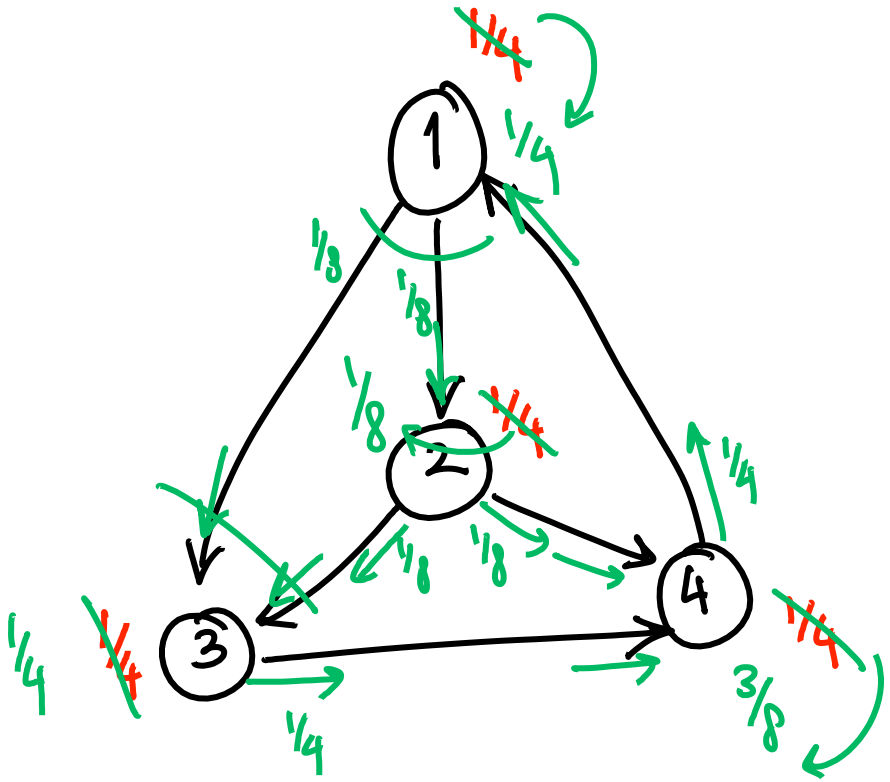
# *PageRank*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

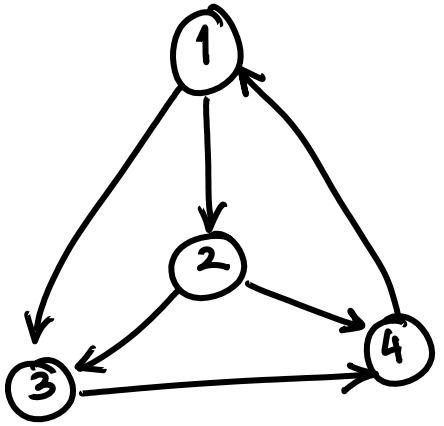
Initialize all pages to have equal PageRank 1/N (where N = total number of pages)

- Repeat forever
  - Each page divides its PageRank equally among its out-links (if none, retain all of it)
  - Each page updates to the sum of PageRanks shares received from all in-links
- Converges to a limiting value

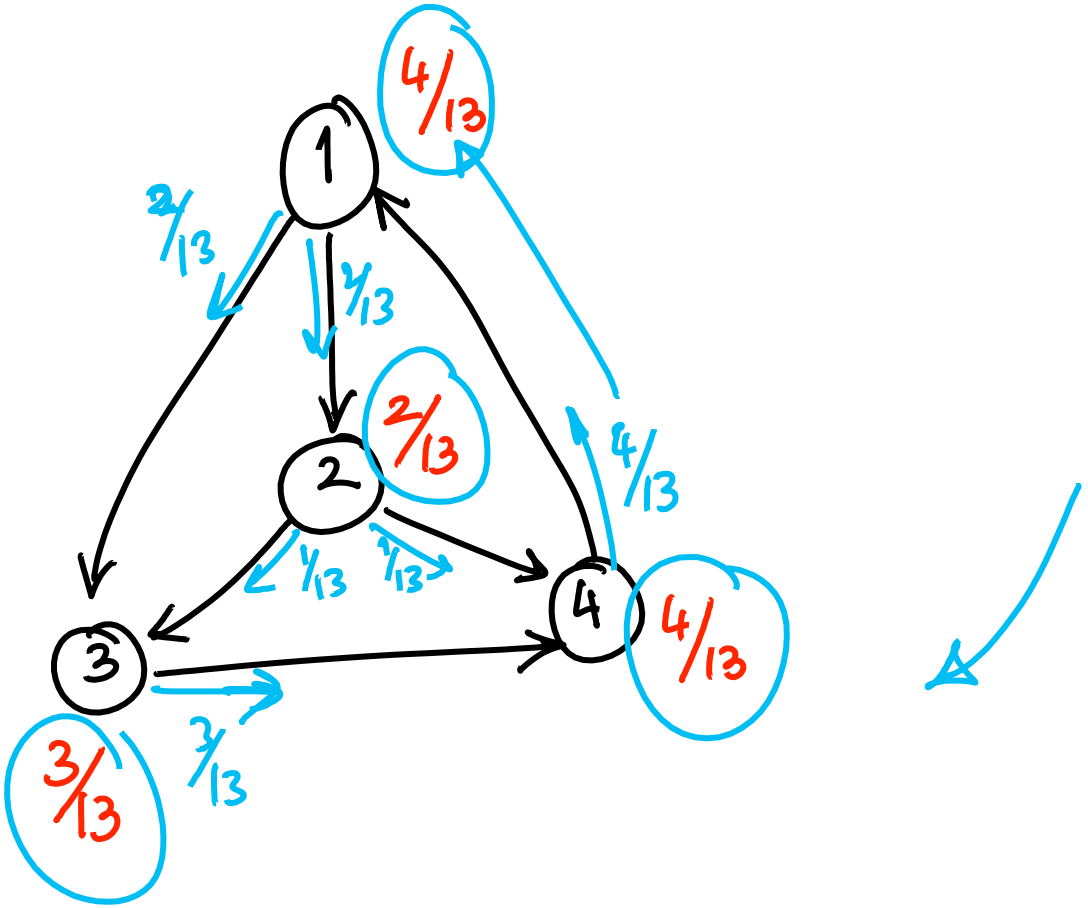
# Example



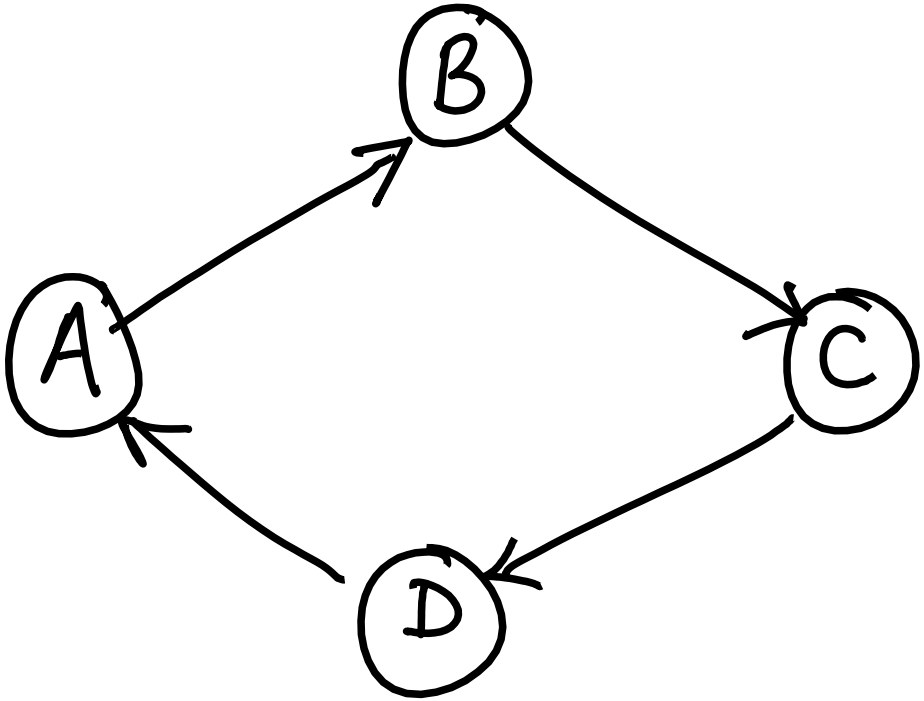
# *Spreadsheet setup?*



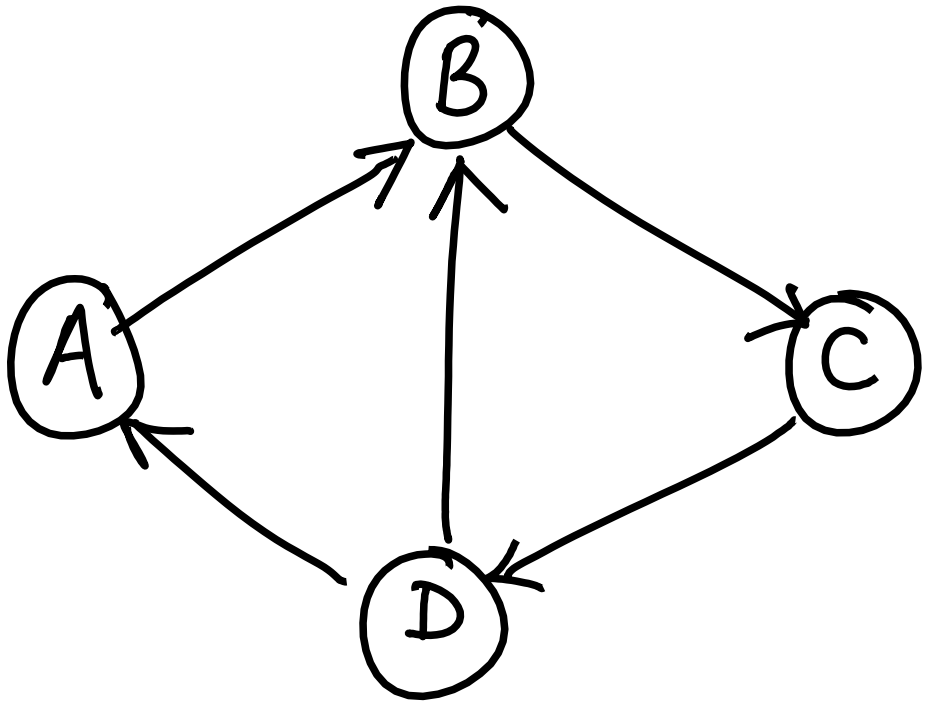
# *Example: Converged values*



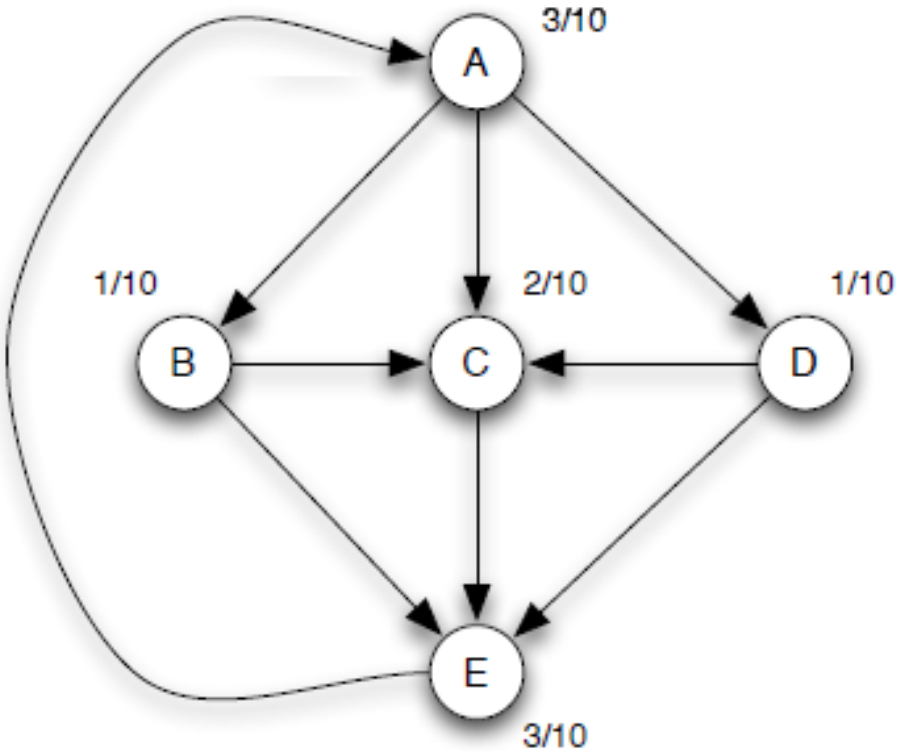
# Another Example



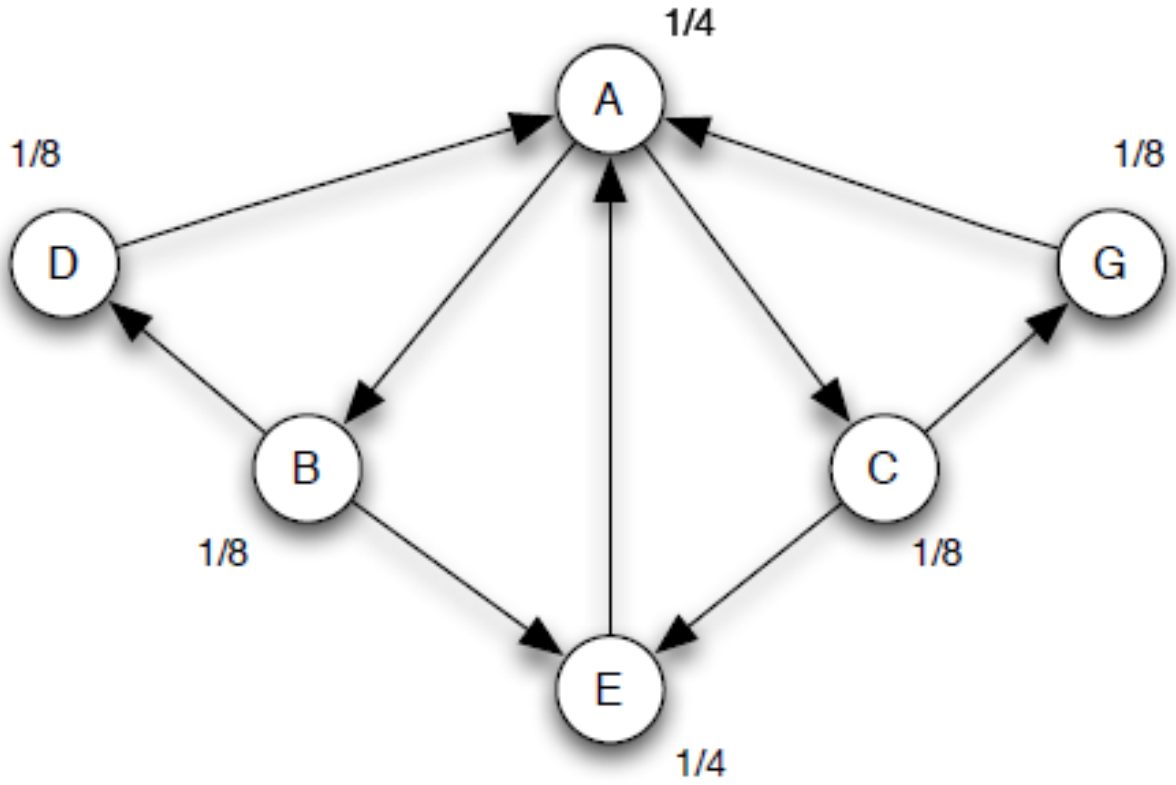
# Another example in the video



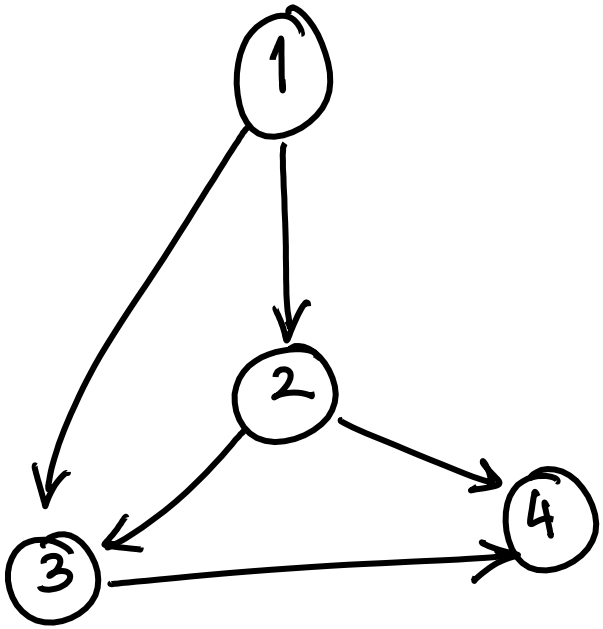
# *Are these values at equilibrium?*



# *Are these values at equilibrium?*



# Difficulty



# *Scaled PageRank*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

Pick scaling factor $0 < s < 1$ and shrink total PageRank in graph from 1 to s

Divide residual (1-s) equally among all nodes in ***every*** update

In practice, $0.8 < s < 0.9$

# *Random Surfer Model*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

Another equivalent view of scaled PageRank

- Random surfer goes to a random page with probability (1-s)
- Otherwise, chooses one outgoing link and clicks through to it
- PageRank of a page is the chance of finding surfer there in the long run

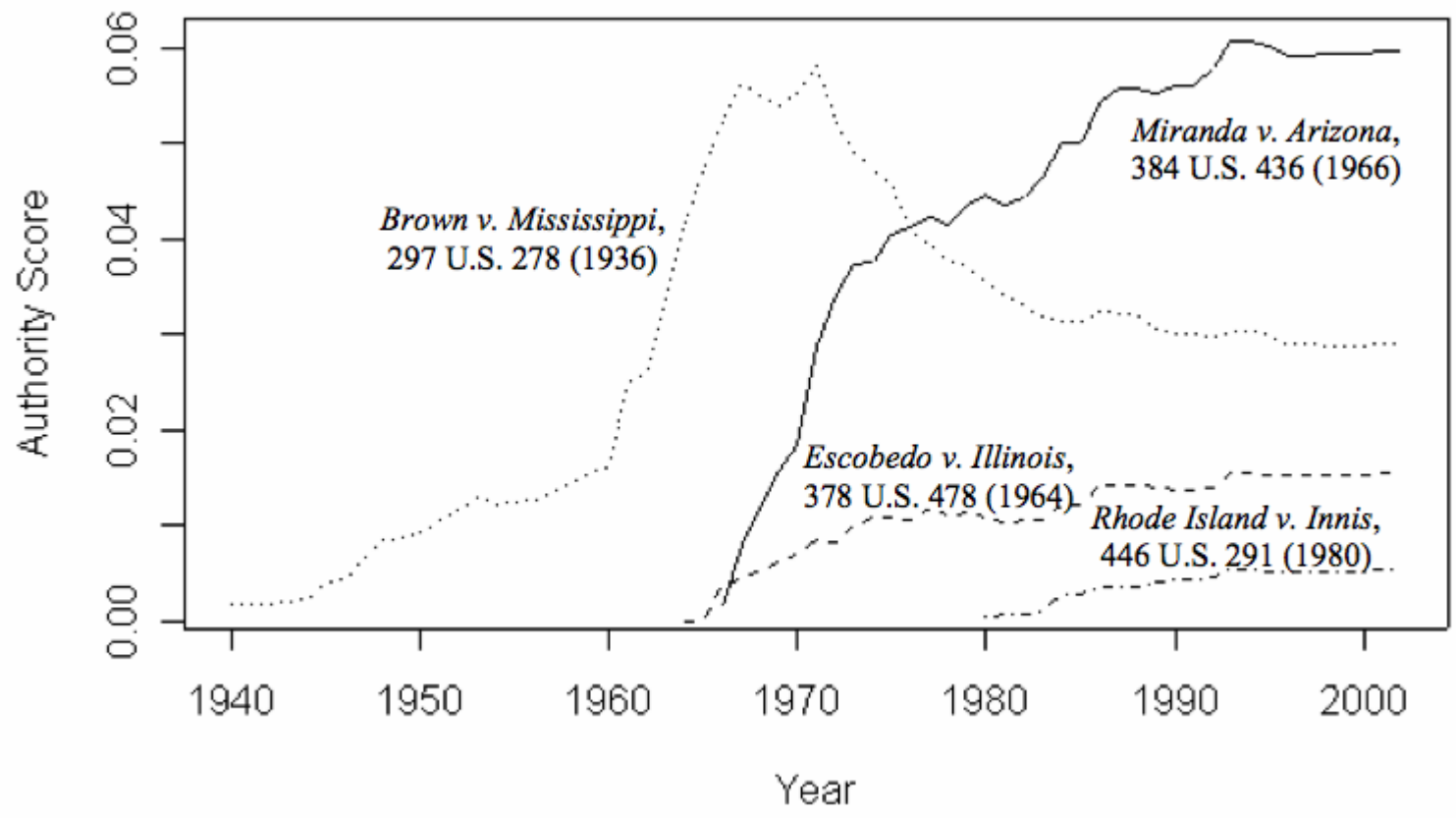
# *Further Refinements*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- Anchortext
- User feedback
- Search Engine Optimization (SEO) → Need to keep ranking method a moving target

# *Citation Analysis*

Carnegie Mellon Tepper SCHOOL OF BUSINESS

- Impact factor for journals [Garfield]
  - Citations received in past two years
- Influence weights [Pinski & Narin]
  - Notion similar to PageRank for journals

# *Authority scores over time*