Ch. 14: Link Analysis and Web Search

Bishesh Khadka

Search “MIT” ---> How do we get to www.mit.edu?

Problem of Ranking

- Search is hard
  - Information retrieval systems
  - Keywords
    - Synonymy
    - Polysemy
- Ranking on web is harder
  - Abundance of information
  - Content credibility

Link Analysis: Voting by In-Links

- No intrinsic “rank” value in web pages
- Aggregate the number of In-Links
    - In-Links = Endorsements
- Algorithm:
    1. Find “sample” of “relevant” pages
    2. Aggregate In-Links
    3. Rank based on In-Link counts

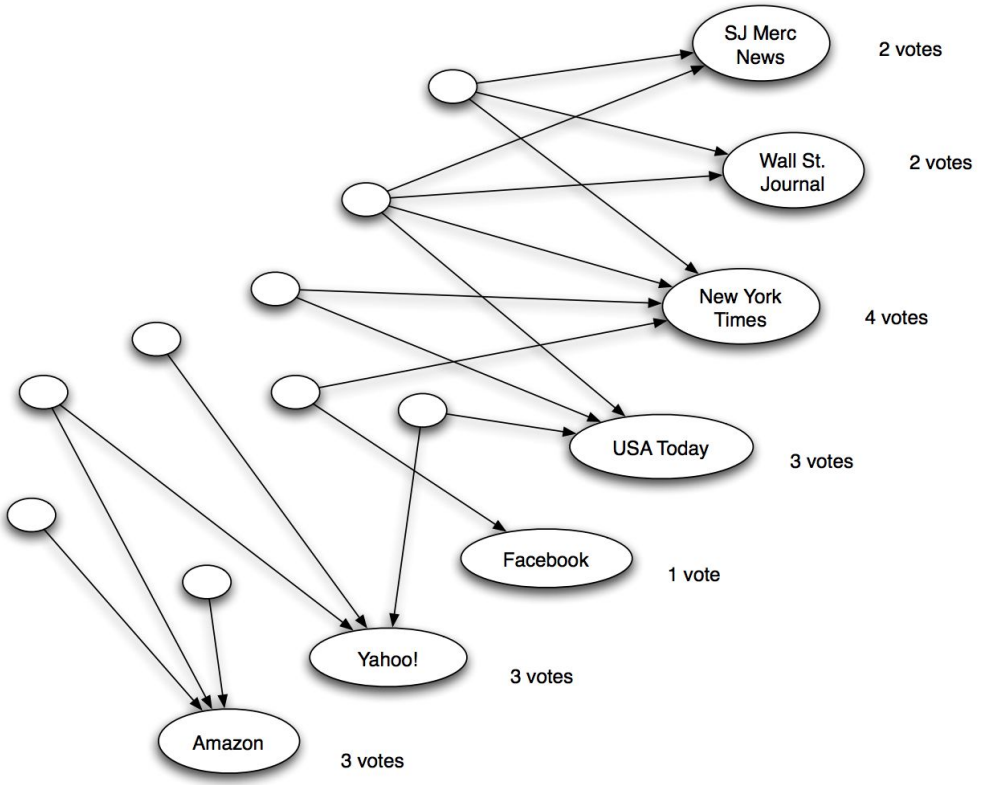


Figure 14.1: Counting in-links to pages for the query "newspapers."

Link Analysis: List-Finding Technique

- In-Link voting isn’t perfect
    - Skewed to pages with most In-Links
    - Even irrelevant ones

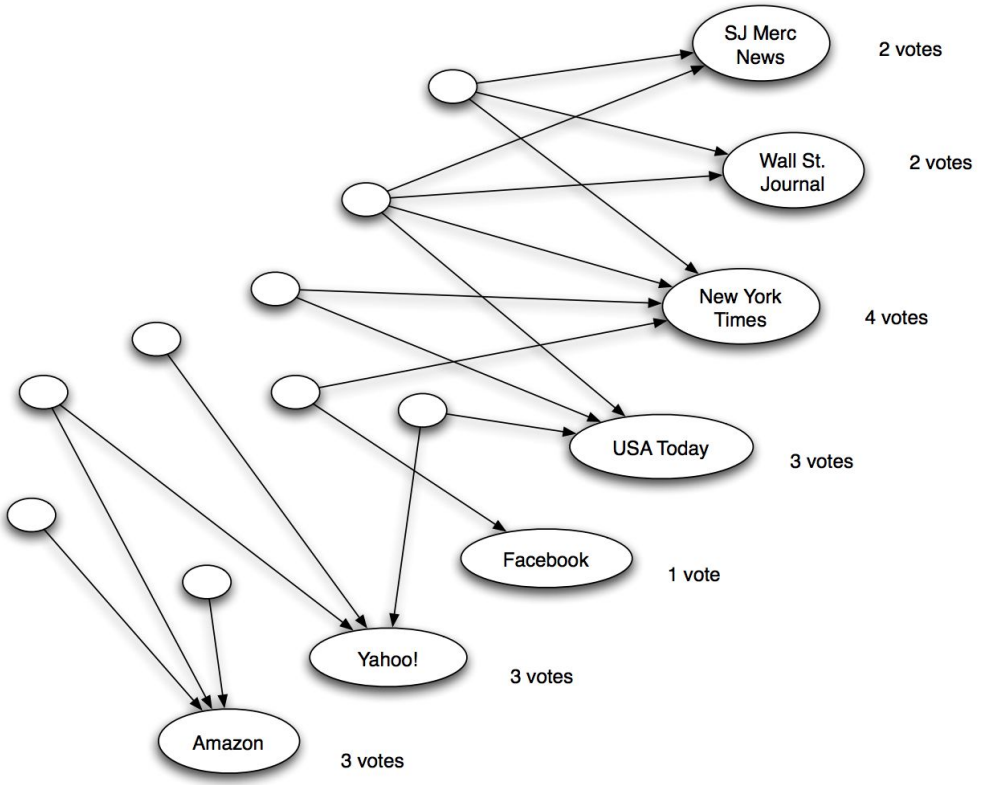


Figure 14.1: Counting in-links to pages for the query "newspapers."

Link Analysis: List-Finding Technique

- In-Link voting isn’t perfect
    - Skewed to pages with most In-Links
    - Even irrelevant ones
- “Hub” pages
- “score” is sum of votes for pages it points to

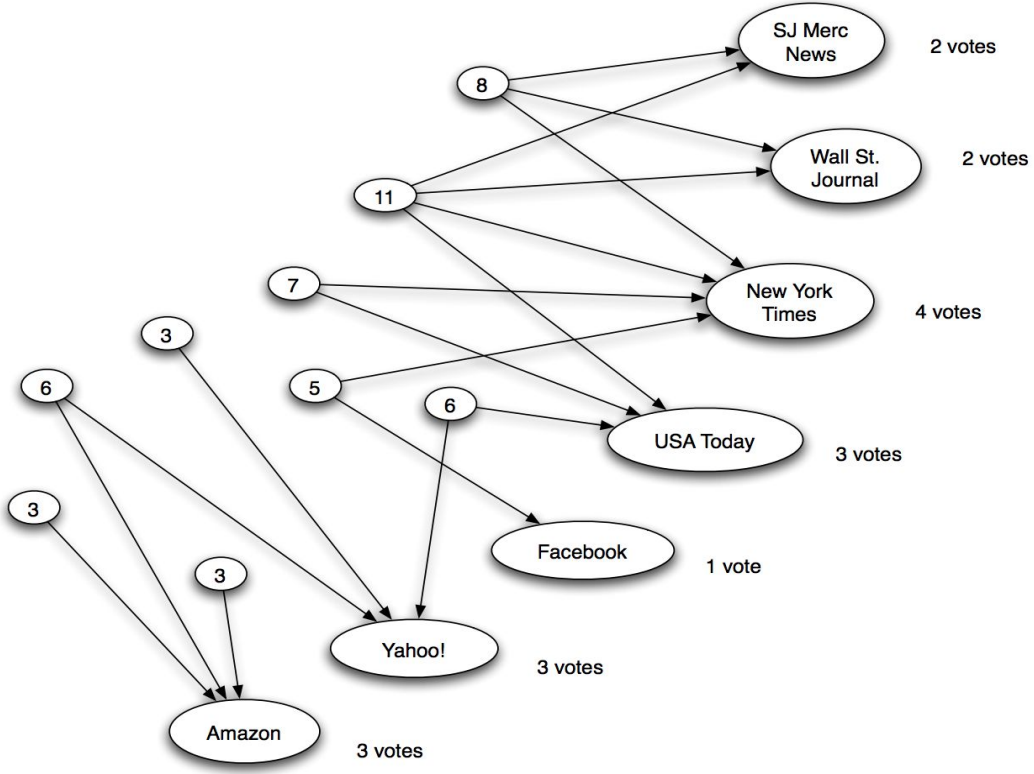


Figure 14.2: Finding good lists for the query "newspapers": each page's value as a list is written as a number inside it.

Link Analysis: Repeated Improvement

- Intuition: Lists with links to “good” sites are credible
- Pages with list compilations are “hubs”
- Pages these hubs point to are “authorities”
- Algorithm:
  1. All hubs and auths have score 1
  2. For k iterations:
     - $\forall$ auth page p: auth(p) = $\sum$ hub(j) $\forall$ j hubs that have voted for p
     - $\forall$ hub page p: hub(p) = $\sum$ auth(j) $\forall$ j auths that p has voted for

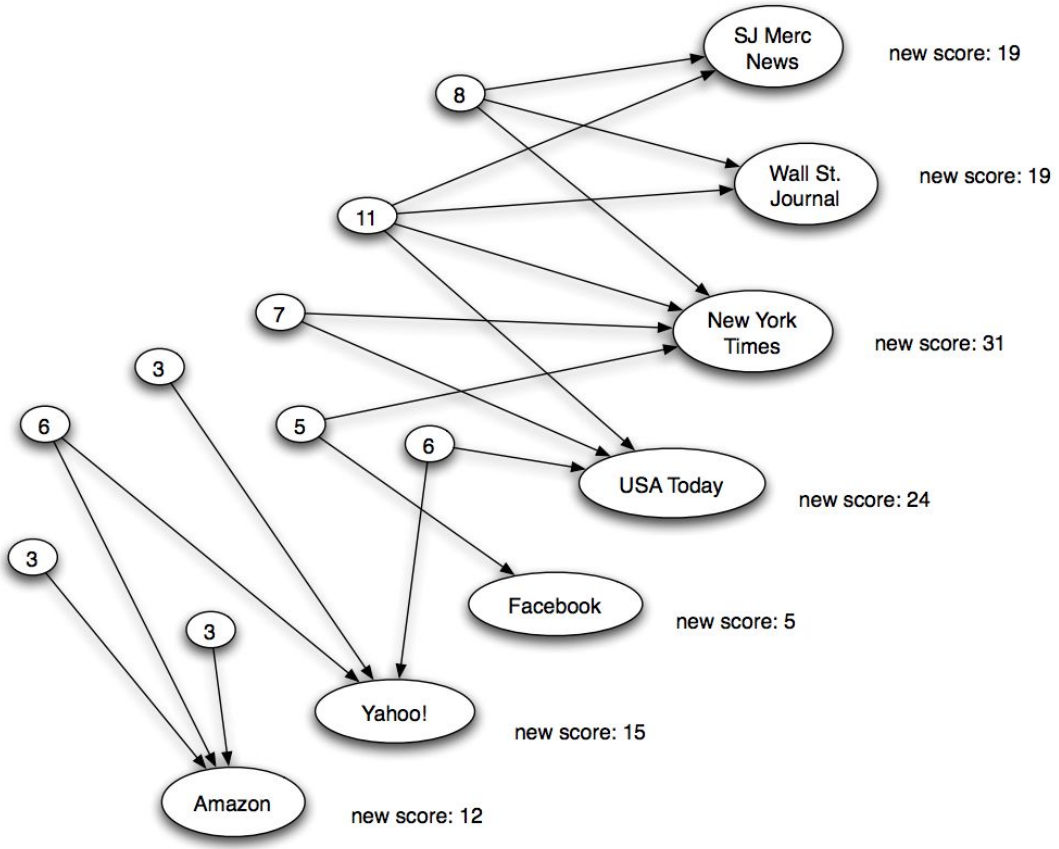


Figure 14.3: Re-weighting votes for the query "newspapers": each of the labeled page's new score is equal to the sum of the values of all lists that point to it.

Link Analysis: Repeated Improvement

- Hub and auth scores normalized between each set of pages
- Scores stabilize as k gets large

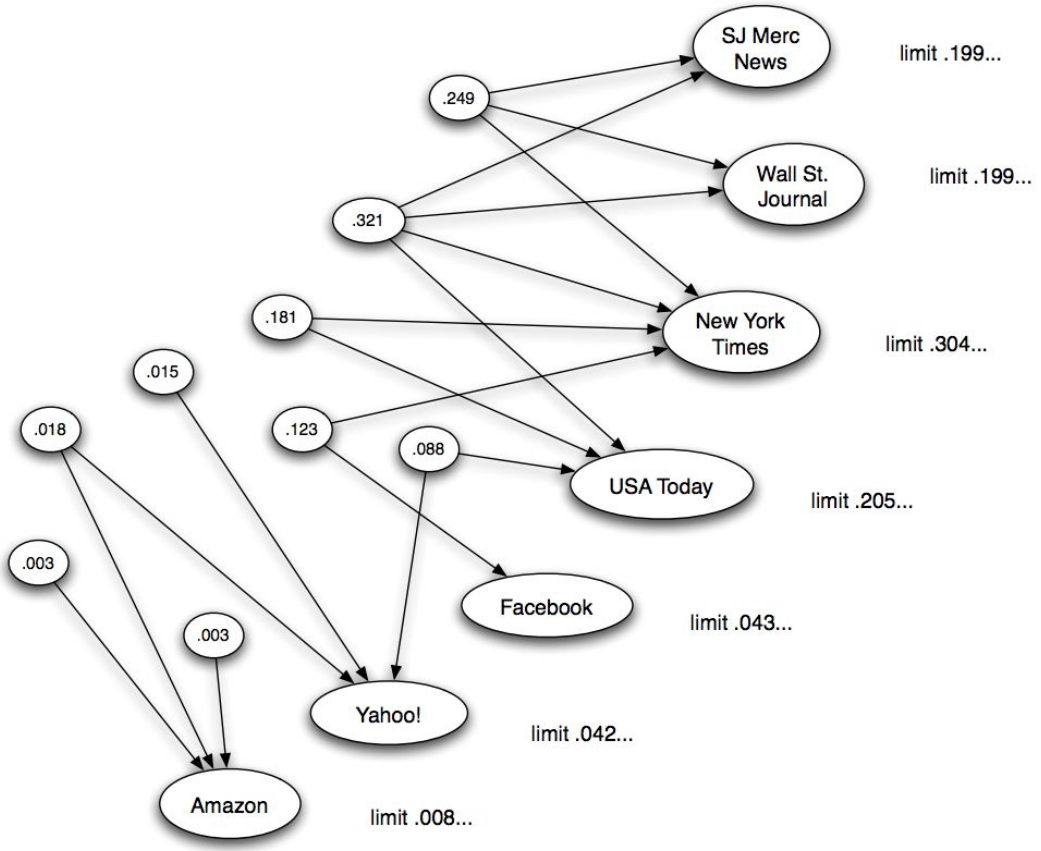
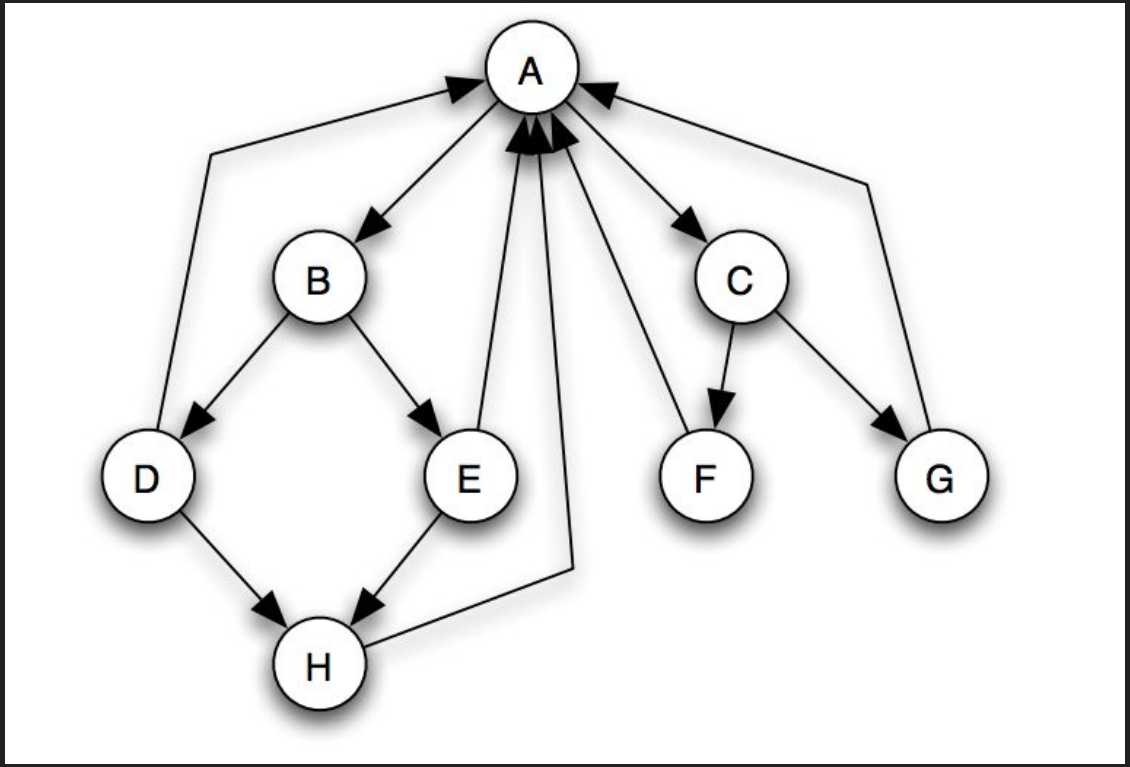


Figure 14.5: Limiting hub and authority values for the query "newspapers."

PageRank

- Intuition: a page is important if it is cited by other important pages
- Algorithm:
  1. $\forall$ page i $PageRank_i = 1$
  2. For k iterations:
     - $\forall$ page i send $PageRank_i$ / (# outgoing edges in i) to every outgoing edge
     - Update all PageRank values to be $\sum$ received



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

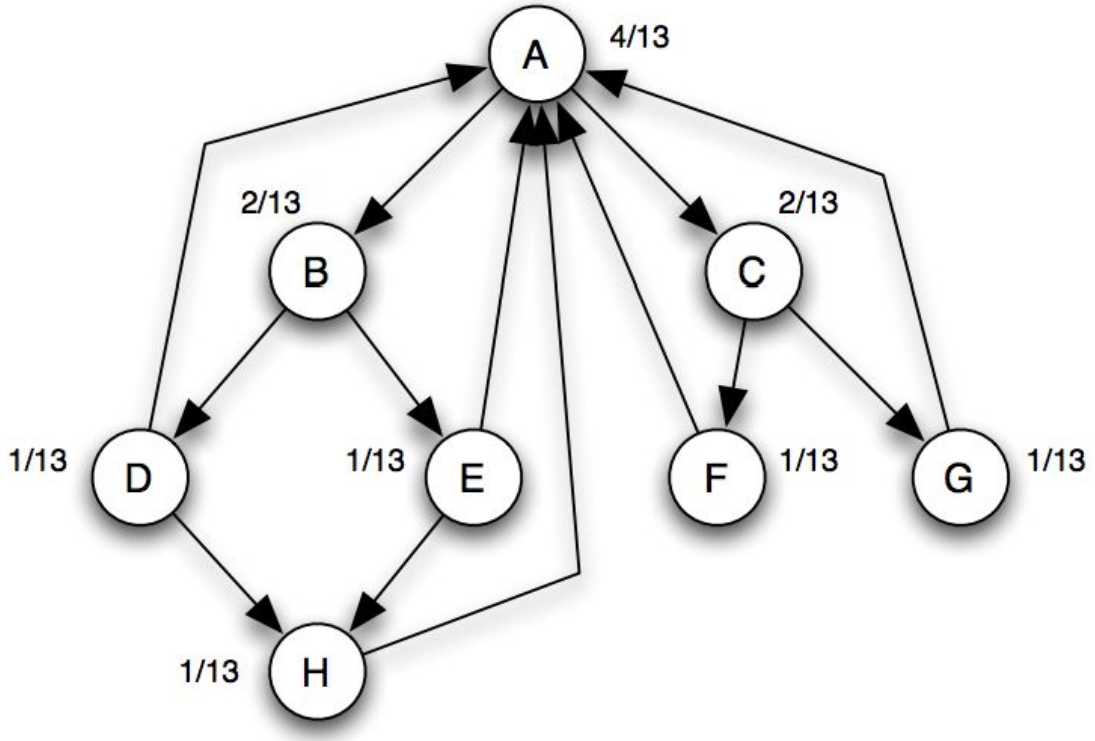


Figure 14.7: Equilibrium PageRank values for the network of eight Web pages from Figure 14.6.

PageRank: Scaled

- Invalid nodes can end up with all the PageRank

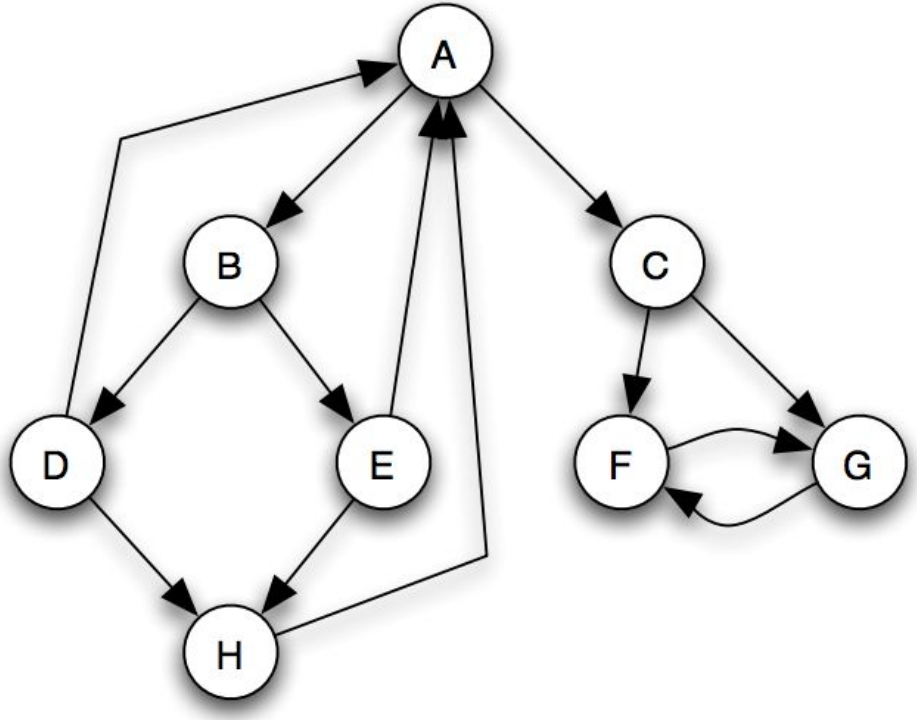


Figure 14.8: The same collection of eight pages, but F and G have changed their links to point to each other instead of to A. Without a smoothing effect, all the PageRank would go to F and G.

PageRank: Scaled

- Invalid nodes can end up with all the PageRank
- Intuition: all water going to deepest point
- Scaled Algorithm:
  1. $\forall$ page i $PageRank_i = 1$
  2. For k iterations:
     - Perform normal PageRank updates
     - Scale all PageRanks by factor s
     - Add (1-s)/n PageRanks to all nodes

S = 0.8 - 0.9 in practice

PageRank: Random Walk Definition

- The probability of being at a page X after k steps of random walk is precisely the PageRank of X after k applications of the Basic PageRank Update Rule
- Scaled: with probability s the traveler follows random edge as before, but with probability 1 - s the traveler jumps to any random node
- Proof in 14.6

Link Analysis: Beyond Web

- Authority in network structures
- Publications
- Supreme Court Cases

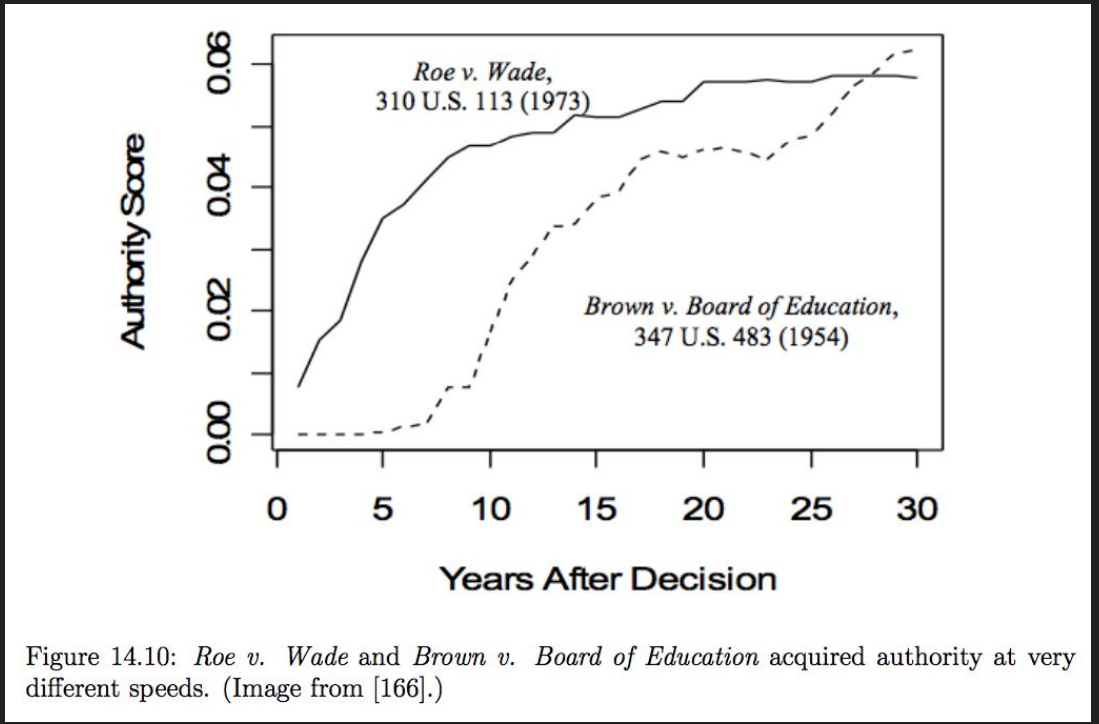


Figure 14.10: *Roe v. Wade* and *Brown v. Board of Education* acquired authority at very different speeds. (Image from [166].)

- References
  - D. Easley, J. Kleinberg, Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Cambridge University Press, Cambridge, UK, 2010