CSC200: Lecture 24

Allan Borodin

Announcements and this lecture

- I am adding a question to the current assignment 3. I will try to add relevant questions as topics are being disucssed. I advise keeping up with assignments and not waiting until close to the due date.

- This lecture Search engines (Ch 14)

Chapter 14: Link Analysis and Web Search

- We really should be rather impressed if not surprised by how good (essentially key word) search engines (e.g. Google) seem to be in most cases.

- Basic IR (information retrieval) problem: take a (complex or ambiguous) information need (expressed by just keywords) as input and produce a ranked list of relevant documents.

- 1960s-70s (maybe even in the 80s) debate:
  - Was computerized search mainly an algorithmic problem (like other say optimization problems), or
  - intrinsically a problem of artificial intelligence requiring us to understand and mimic human intelligence to comprehend meaning?

Chapter 14: Link Analysis and Web Search

- We really should be rather impressed if not surprised by how good (essentially key word) search engines (e.g. Google) seem to be in most cases.

- Basic IR (information retrieval) problem: take a (complex or ambiguous) information need (expressed by just keywords) as input and produce a ranked list of relevant documents.

- 1960s-70s (maybe even in the 80s) debate:
  - Was computerized search mainly an algorithmic problem (like other say optimization problems), or
  - intrinsically a problem of artificial intelligence requiring us to understand and mimic human intelligence to comprehend meaning?
  - Who won the debate?

Why is search a difficult problem?

Many issues making search and especially keyword search difficult:

- Synonymy: many different words for same concept; vertex/node, equal/same, etc.
- Polysemy: same word having many meanings; e.g. jaguar, cougar/puma
- Abundance of writing styles (everyone is an author) and abundance of information needs (everyone searches)
- Somewhat static data (the web is not crawled everyday) vs constantly changing events
- Scarcity ("needle in a haystack" requests was the norm in IR) vs todays predominance of abundance of relevant responses.
  - This abundance requires search engines to rank documents.
  - The importance of the "top 10".

Content, links, usage

- Given the emphasis of the text being networks, Ch 14 concentrates on how link analysis contributes to the ranking of documents.
- As indicated in Section 14.4, search engine ranking utilizes sophisticated combinations of content, link and usage data.
- Moreover, search engine specifics necessarily change over time as there is a game-theoretic aspect of search engine companies vs. web page optimizers.
- Knowing precisely how a search engine chooses and ranks documents in response to a query, one can modify a document to score highly (even if document not at all relevant to the query).
- "IR has become statistical machine learning"
- Indeed one can argue that much of AI has become statistical machine learning

Primitive view of content analysis

The most basic approach to using the syntactic content of a document (i.e. the words and groups of words in the document) to identify plausibly relevant documents:

- Ignoring synonymy and polysemy, let's consider a document as a bag of words
- Additionally can treat some common word pairs such as "mixed bag", "first responders" and perhaps even triples of common word sequences "world wide web" as if they were single words).
- Can collect all document identifiers containing a particular word in a sorted list
- Then given a query (which is usually 2 or 3 words), we can quickly find all the documents containing all or most of the query words.

A few other content ideas

- Can use word counts especially when relativized to overall frequency occurrence of words. Use of "td-idf" meaning *term frequency/inverse document frequency*
- Can deal to some extent with polysemy by context; for example, occurrence of other (especially nearby) words
- Can use common synonyms
- Can emphasize words that occur early in the document, in titles, in section headings, in anchor text.
- Example of hyperlink and anchor text:

```
<a href="http://www.cs.toronto.edu/~bor/"> Allan's
website </a>
```

A few other content ideas

- Can use word counts especially when relativized to overall frequency occurrence of words. Use of "td-idf" meaning *term frequency/inverse document frequency*
- Can deal to some extent with polysemy by context; for example, occurrence of other (especially nearby) words
- Can use common synonyms
- Can emphasize words that occur early in the document, in titles, in section headings, in anchor text.
- Example of hyperlink and anchor text:
  `<a href="http://www.cs.toronto.edu/~bor/"> Allan's website </a>`

Just in case you want to link to my home page.

Link analysis

Jon Kleinberg

Larry Page and Sergey Brin

- Although there are differences of opinion as to the current relative significance of link analysis in search ranking, it initially played a very significant role and it continues to play a role in ranking. (See §14.4)
- Kleinberg (IBM) and Brin & Page (at Stanford/Google) were independently developing algorithms for exploiting the hyperlinks in web pages.
- Kleinberg's "Hubs and Authorities" was implemented in a prototype (not publicly available) system at IBM; Brin and Page implemented "Page Rank" (named for Page) in Google.
- Both approaches can be best understood in terms of linear algebra and in terms of eigenvalues and eigenvectors (spectral analysis).

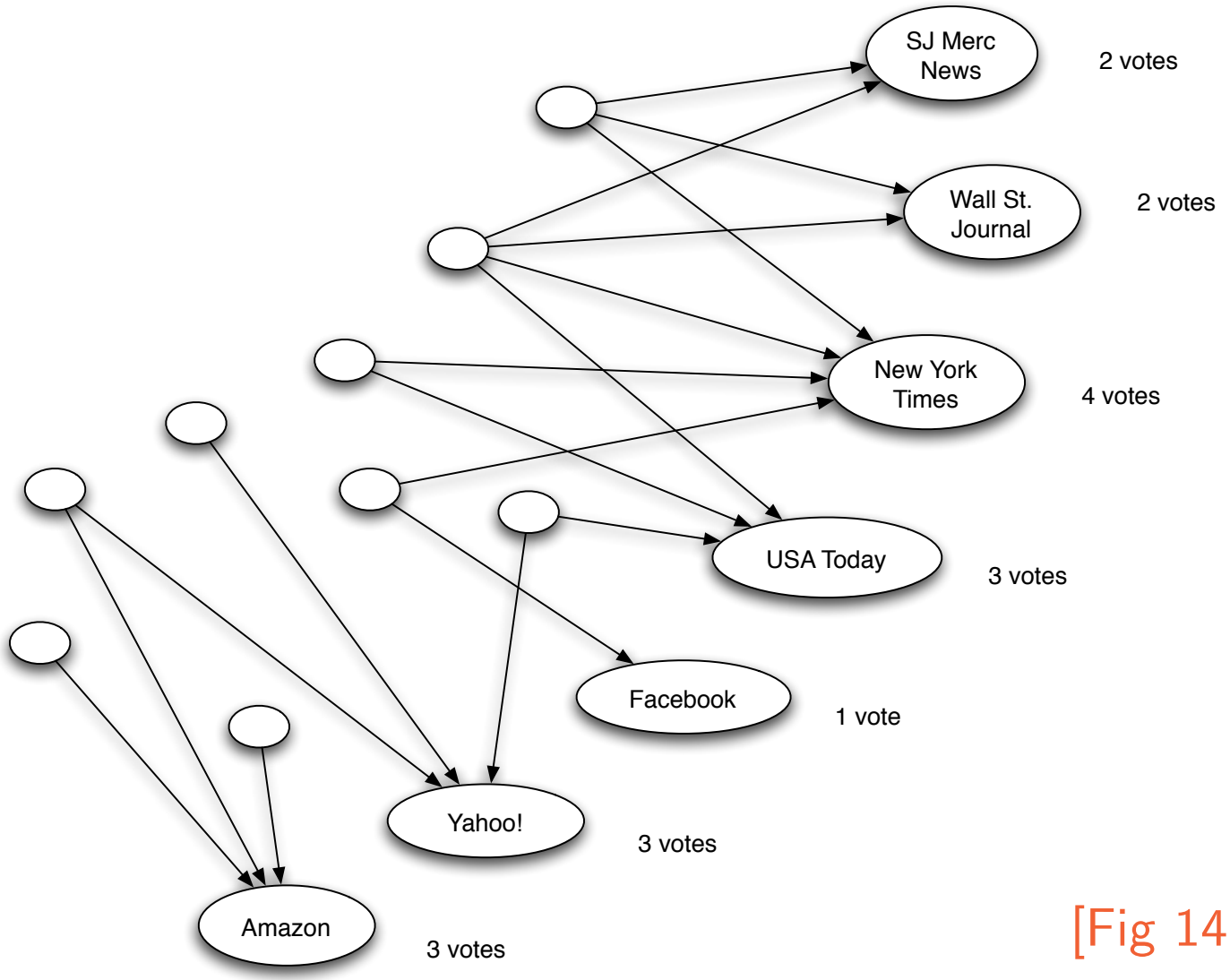
Hubs and Authorities

- The simplest way to utilize links to rank web pages would be to think of each link from A to B as an endorsement or vote by A for B.
- And then use the number of endorsements as a key feature determining the rank. Of course, one would have to adjust such scores coming from say the same domain name.
- Even after adjusting for such "vote fixing", if Dion Phaneuf has a web site and a link suggesting where he buys his hockey equipment you might think that is more meaningful than say where I buy hockey equipment (especially since I don't play hockey).

Reinforcement of Hubs and Authorities.

- This then becomes the motivation (and seemingly circular reasoning) behind hubs and authorities.
- The best "authorities" on a subject (places to buy equipment) are being endorsed by the best hubs (people who know where to buy equipment).
- Similarly, the best hubs are those sites that recommend the best authorities. Conceptually consider the link structure as setting up a bipartite graph. The same web page can be both a hub and an authority.
- This idea is nicely explained in Figures 14.1-14.5 of E&K.

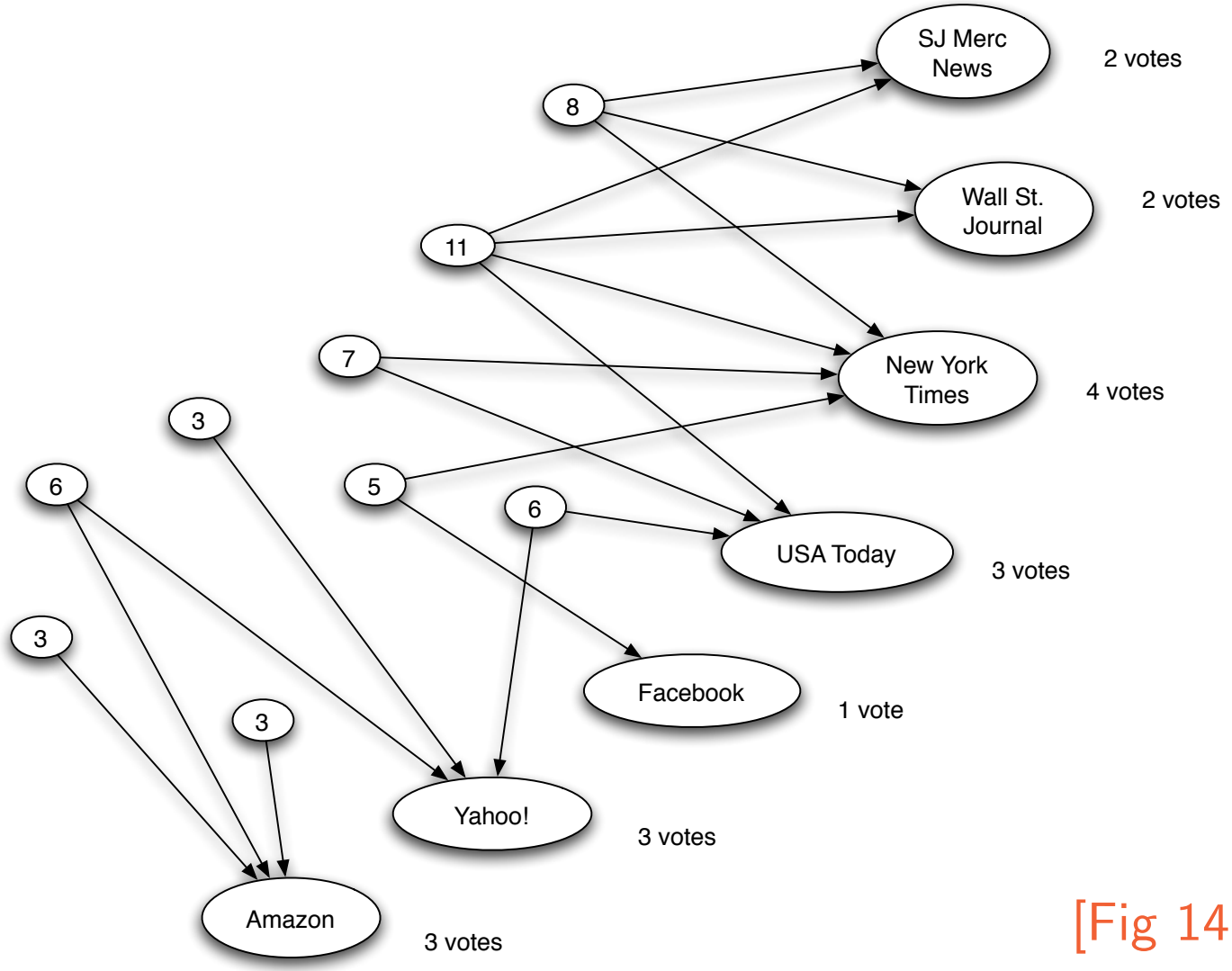
- The result of applying the authority update rule: for each page p, auth(p) is the sum of hub values (initially just the number) of hubs pointing to p.



[Fig 14.1, E&K]

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

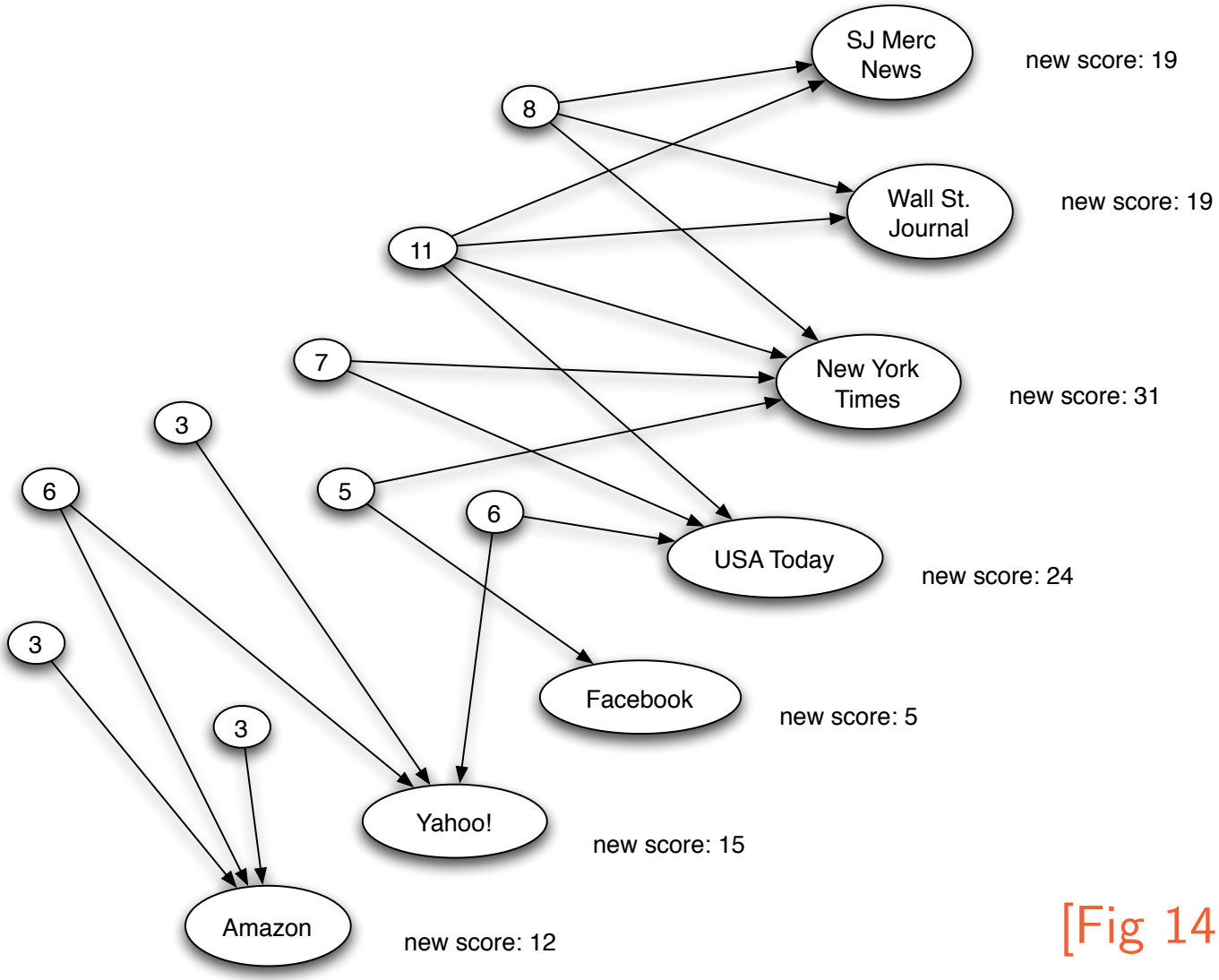
- Then to recalibrate hub values, we use the hub update rule: for each page p, hub(p) is the sum of values of all authorities that p points to.



[Fig 14.2, E&K]

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

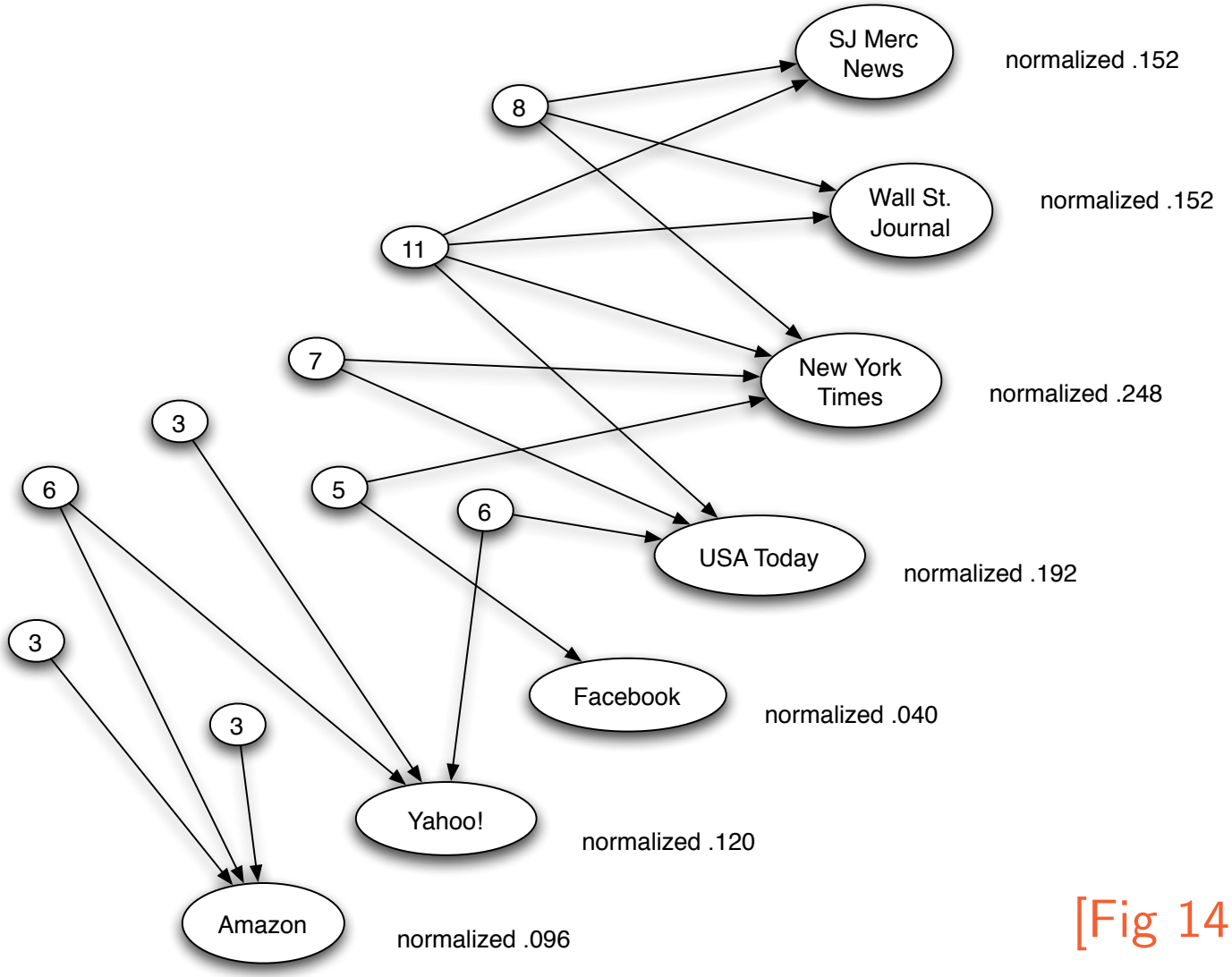
- Applying the authority update rule again we get figure 14.3.



[Fig 14.3, E&K]

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

- Since we only care about the relative values of these numbers, both authority and hub scores can be normalized to sum to 1 (to avoid dealing with large numbers).



[Fig 14.4, E&K]

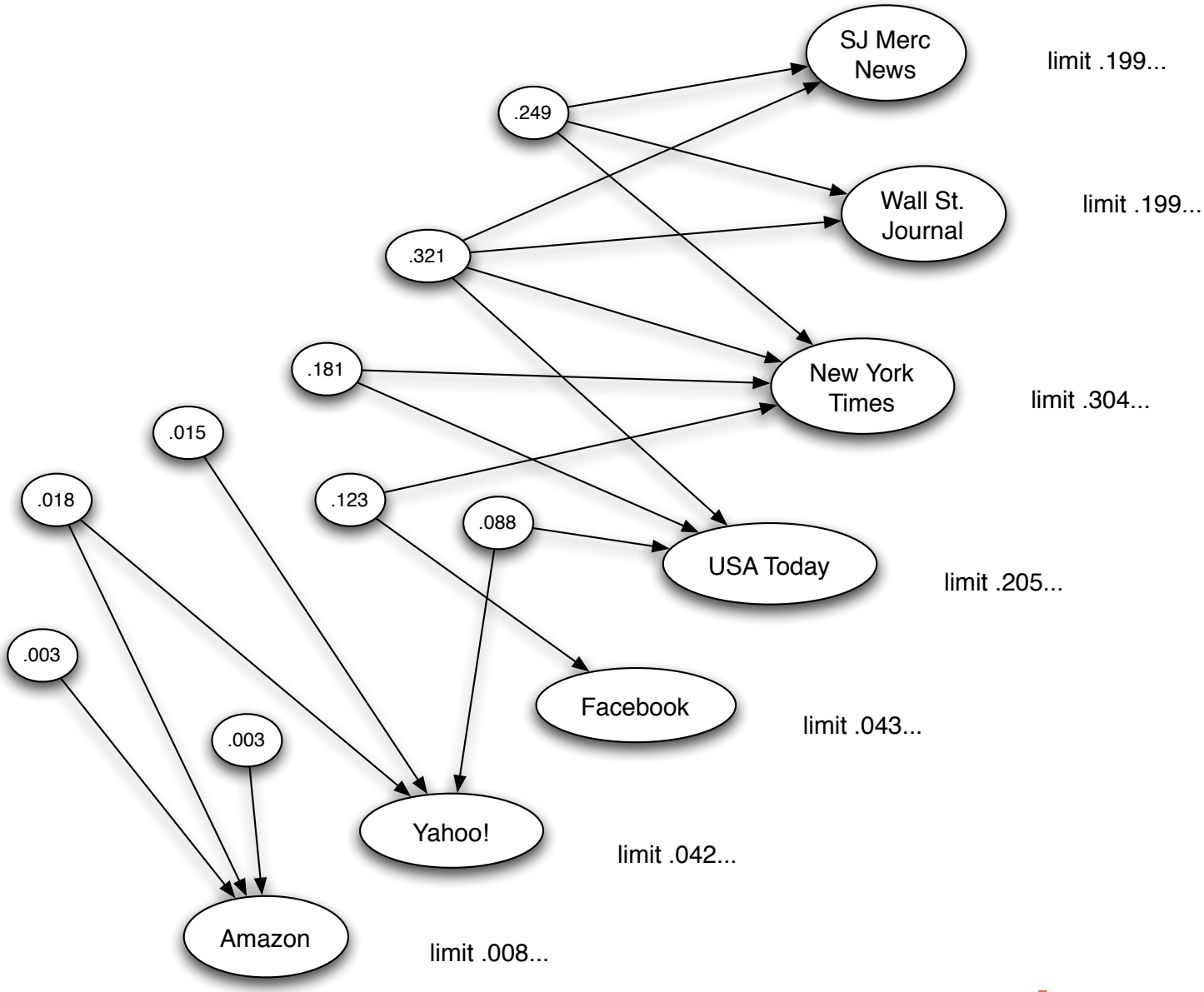
Figure : Re-weighting votes after normalizing for the query "newspapers".

Keep repeating a good idea

- Now having recalibrated and normalized both the authority and hub scores, we can continue this process to continue to refine these scores.

- That is, the hubs and authorities procedure is as follows:
  - Initialize all hub values (say to some positive vector perhaps depending on usage or content)
  - For sufficiently large k, perform the following k times
    - ★ Apply authority update rule to each page
    - ★ Apply hub update rule to each page
    - ★ Normalize so that sum of A and H weights $= 1$.

- Using linear algebra, it can be shown (in Section 4.6) that these A and H normalized values will converge to a limit as $k \to \infty$ (which can be approximated by some sufficiently large k)!



[Fig 14.5, E&K]

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

Page Rank

- The motivation behind page rank is a somewhat different view of how authority is conferred.
  - Endorsement of authority is conveyed by other authorities
  - (i.e. no hub concept).
  - This is how peer review works in the academic and scholarly world.

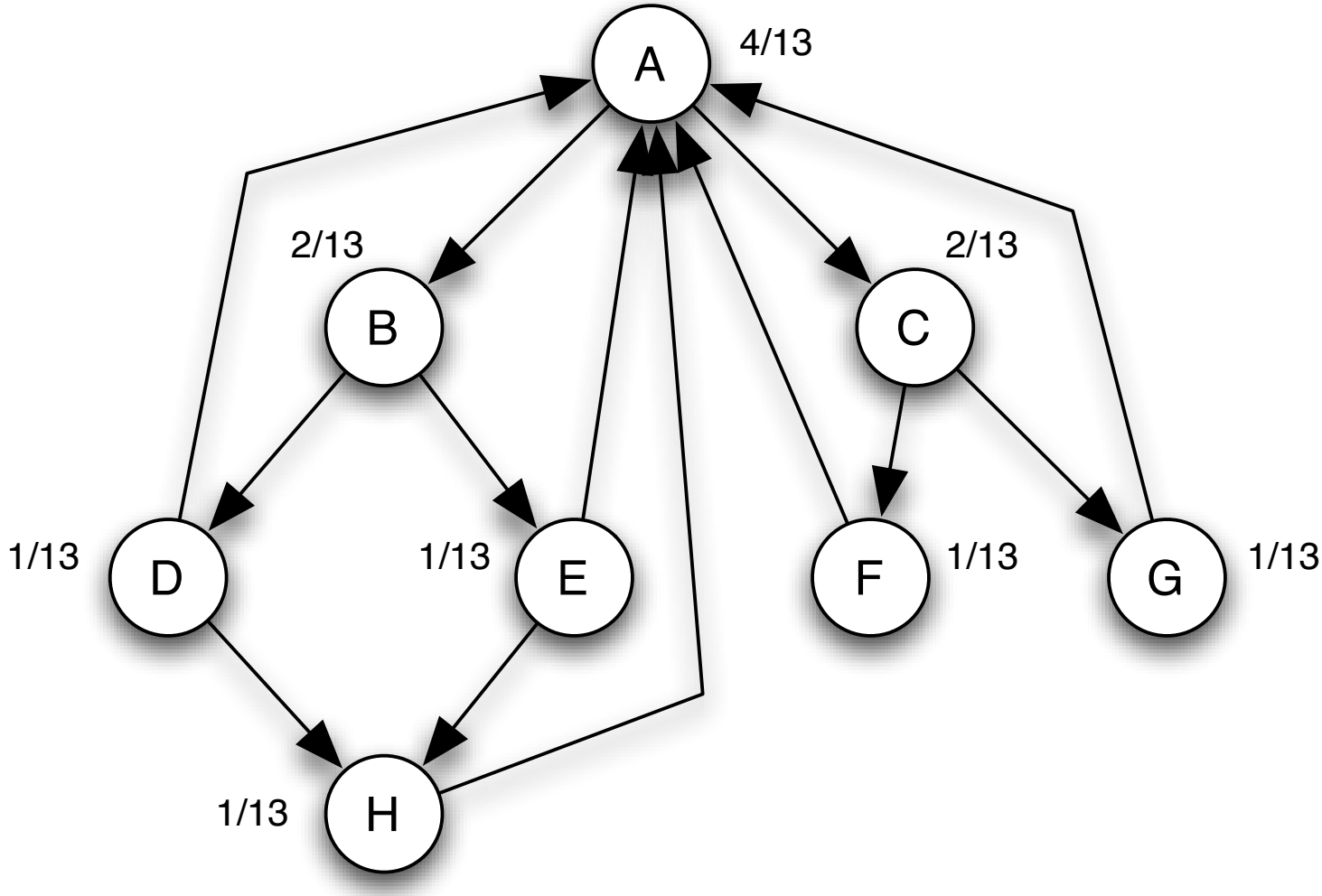
- Two equivalent views of page rank:
  1. More directly models this idea of authorities conveying authority.
  2. Reformulates this in terms of a random walk on a graph.

Keep repeating a good idea

- Suppose at any point of time we have authority scores for all relevant pages.
  - A page spreads its authority equally amongst all of its out links.
  - If a page has no outlinks then all authority stays there.
- This redistributes the authority scores. (We are not creating or losing any authority, we are just redistributing it.)
- We can initially start with every relevant page having authority $1/n$ where there are n pages. Then we repeat this process k times for some sufficiently large k .
- It can be proven (again using linear algebra) that this process has a limiting behavior as $k \to \infty$.

Remark

In many cases this won't reflect the desired authority. Namely, if the network has any sinks (or SCC that are sinks) which it will surely have, then all of the authority will pass to such sinks.



[Fig 14.7, E&K]

Figure : Equilibrium PageRank values for the network of eight Web page.

Scaled page rank

- The way around this sink hole of authority is to have a scaled version of page rank where
  - only a fraction s of the authority of a page is distributed to its out links
  - the remaining $(1 - s)$ fraction is distributed equally amongst all relevant pages.

- For any value of s, we get convergence to a unique set of scores for each page and that is its page rank (for that particular value of s). It is reported that Google uses $0.8 \leq s \leq 0.9$.

- (See the footnote on page 410 of E&K as to why in the previous example, nodes F and G will still get most of the authority but that for realistically large networks, the process works well.)

Some additional remarks

- The limiting scores for both the authority and hubs approach and the page rank approach are equilibrium points for an appropriate algebraic process.
- That is, if we actually were in the limiting state, we would be in the equilibrium state. In practical computation, we stop the process when the change in each iteration is sufficiently small.
- We can weight the network edges (say according to some concept of link importance) and apply the same authority and hubs or page rank approach distributing authority in proportion to these weights.