Information Retrieval on the Web

DAPI . Information Description, Storage and Retrieval Course
MIEIC, 2020/21 Edition

Sérgio Nunes
DEI, FEUP, U.Porto

Based on Chapters 19, 20 and 21 from Introduction to Information Retrieval, Manning, C. et al. (2008)

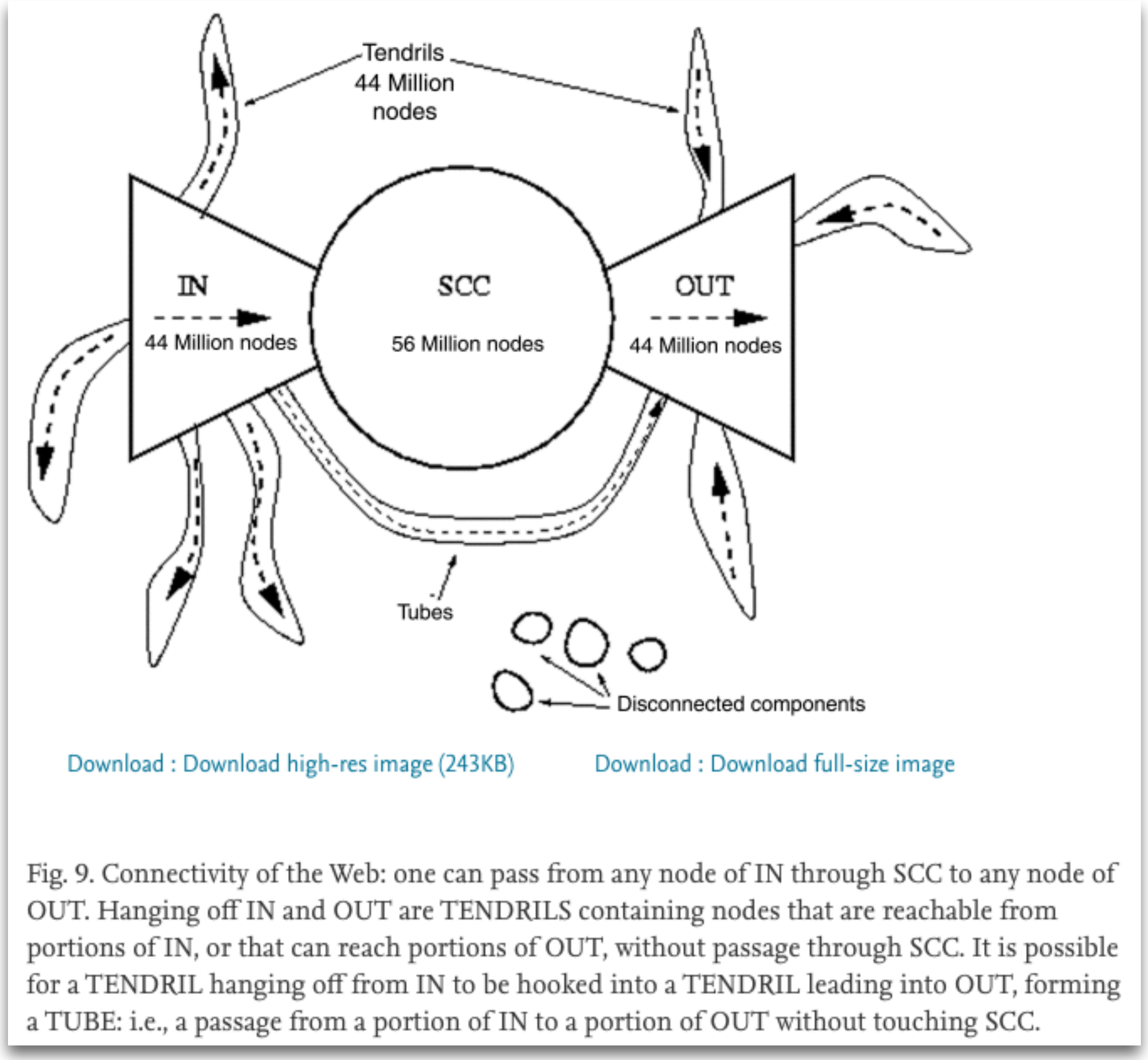
The World Wide Web

- The web is unprecedented in many ways:
  - Unprecedented in scale (size and change);
  - Unprecedented in lack of central coordination;
  - Unprecedented in the diversity of users' backgrounds and needs.
- Two types of challenges:
  - Data: distribution, size, volatility, quality, unstructured, duplicates.
  - Interaction: user needs; relevance; diversity of users.

Web Characteristics

- ➔ The Web can be modeled as a graph.
- ➔ Web pages point to (and are pointed by) other pages.
- ➔ Links to other pages (out-links) usually include an "anchor text".
- ➔ The number of in-links to a page is called the in-degree.
- ➔ Studies of web characteristics and dynamics is an area of research.
- ➔ Early research suggests that the directed graph connecting web pages has a bowtie shape.

Bow-tie Structure of the Web



Download : Download high-res image (243KB)   Download : Download full-size image

Fig. 9. Connectivity of the Web: one can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passage through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE: i.e., a passage from a portion of IN to a portion of OUT without touching SCC.

Web Characteristics

- ➔ There is a (high) commercial value associated with appearing on the top ranked results for a given search.

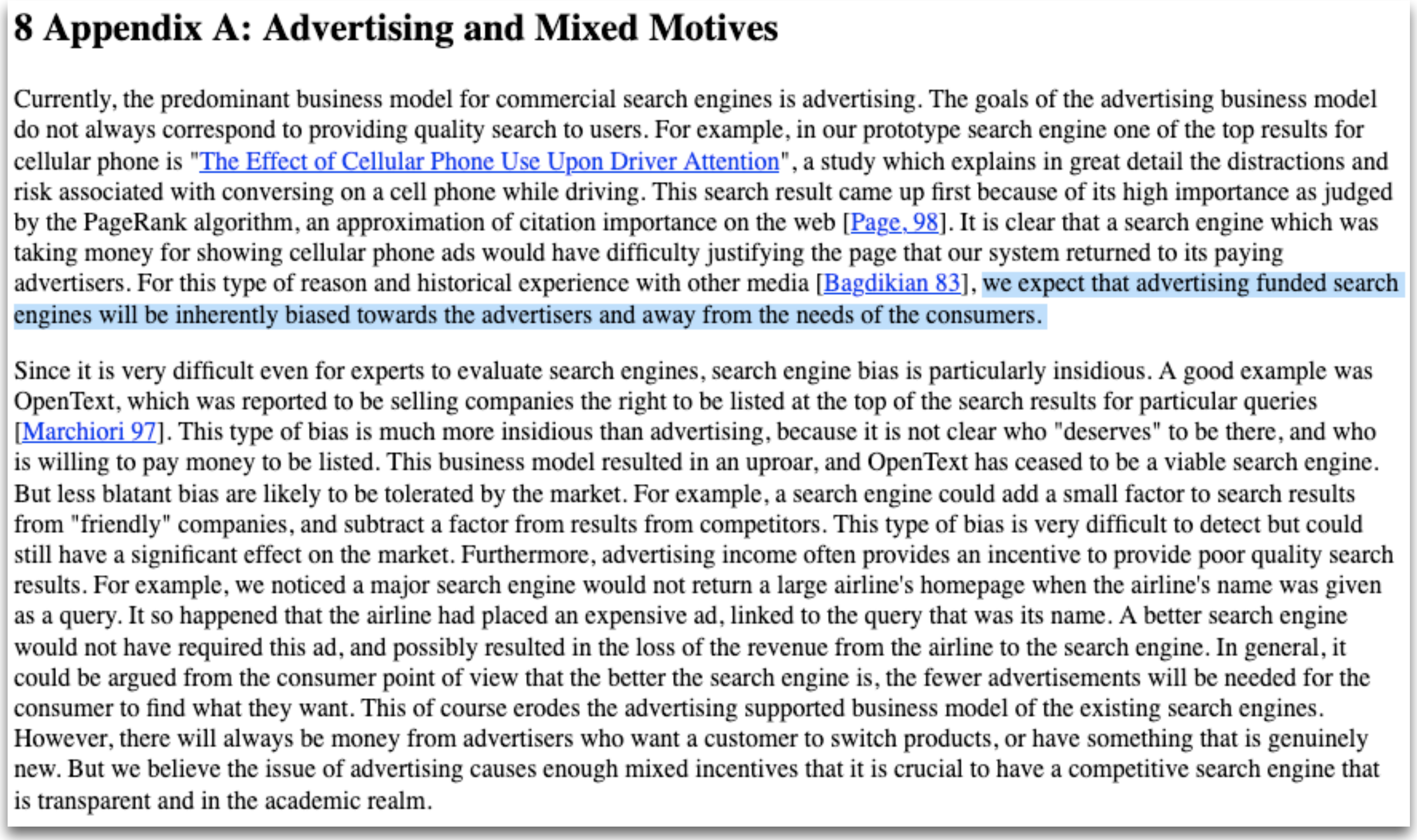
- ➔ For instance, a search engine whose scoring depends on the frequency of keywords would be easy to manipulate by including numerous repetitions of selected keywords.

- ➔ This is called web spam, i.e. the manipulation of content on the web with the purpose of manipulating search engine rankings. Examples include: cloaking, link farms, link spam, click spam, etc.

- ➔ Topics in the sub-area of adversarial information retrieval.

"Mixed Motives in Search"

➔ Reflexion from Brin and Page in 1998.



8 Appendix A: Advertising and Mixed Motives

Currently, the predominant business model for commercial search engines is advertising. The goals of the advertising business model do not always correspond to providing quality search to users. For example, in our prototype search engine one of the top results for cellular phone is "The Effect of Cellular Phone Use Upon Driver Attention", a study which explains in great detail the distractions and risk associated with conversing on a cell phone while driving. This search result came up first because of its high importance as judged by the PageRank algorithm, an approximation of citation importance on the web [Page, 98]. It is clear that a search engine which was taking money for showing cellular phone ads would have difficulty justifying the page that our system returned to its paying advertisers. For this type of reason and historical experience with other media [Bagdikian 83], we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers.

Since it is very difficult even for experts to evaluate search engines, search engine bias is particularly insidious. A good example was OpenText, which was reported to be selling companies the right to be listed at the top of the search results for particular queries [Marchiori 97]. This type of bias is much more insidious than advertising, because it is not clear who "deserves" to be there, and who is willing to pay money to be listed. This business model resulted in an uproar, and OpenText has ceased to be a viable search engine. But less blatant bias are likely to be tolerated by the market. For example, a search engine could add a small factor to search results from "friendly" companies, and subtract a factor from results from competitors. This type of bias is very difficult to detect but could still have a significant effect on the market. Furthermore, advertising income often provides an incentive to provide poor quality search results. For example, we noticed a major search engine would not return a large airline's homepage when the airline's name was given as a query. It so happened that the airline had placed an expensive ad, linked to the query that was its name. A better search engine would not have required this ad, and possibly resulted in the loss of the revenue from the airline to the search engine. In general, it could be argued from the consumer point of view that the better the search engine is, the fewer advertisements will be needed for the consumer to find what they want. This of course erodes the advertising supported business model of the existing search engines. However, there will always be money from advertisers who want a customer to switch products, or have something that is genuinely new. But we believe the issue of advertising causes enough mixed incentives that it is crucial to have a competitive search engine that is transparent and in the academic realm.

http://infolab.stanford.edu/~backrub/google.html

User Characteristics

- Early research has shown that user queries can be grouped in:
  - Informational queries, i.e. seek general information about a topic. There is typically not a single source of relevant information. Users typically gather information from multiple web pages
  - Navigational queries, i.e. seek the website or home page of a single entity. Users' expectations is to find a specific resource.
  - Transactional queries, i.e. preludes the user performing a transaction on the web, such as purchasing a product, downloading a file, or doing a reservation.

Signals for Web Ranking

- ➔ Hundreds of signals are used by search engines to estimate quality.
- ➔ Signals can be groups in different dimensions:
  - ➔ Query-independent signals (static).
  - ➔ Query-dependent signals (dynamic).

  - ➔ Document-based signals (content or structural), e.g. HTML.
  - ➔ Collection-based signals, e.g. Links.
  - ➔ User-based signals, e.g. Clicks.

Link-based Signals

- Base assumption: the number of hyperlinks pointing to a page provides a measure of its popularity and quality.
- Link-based ranking algorithms build on the assumption that an hyperlink from page A to page B represents an endorsement of page B, by the creator of page A.
- Two classic algorithms: PageRank and HITS.
  - PageRank, Larry Page and Sergey Brin (1996)
  - HITS, Jon Kleinberg (1997)

PageRank

- ➔ The PageRank of a node is a value between 0 and 1.
- ➔ It is a query-independent value computed offline that only depends on the structure of the web graph.
- ➔ The algorithms models a random surfer who begins at a web page and, at each step, randomly chooses between the out-links to move to the next page. If the random surfer executes this "forever", he will visit some pages more frequently than others. The PageRank value of a page represents this probability.

$$PR(a) = \frac{q}{T} + (1-q)\sum_{i=1}^{n} \frac{PR(p_i)}{L(p_i)}$$

PageRank Example

- ➔ From "The PageRank Citation Ranking: Bringing Order to the Web" (1998).

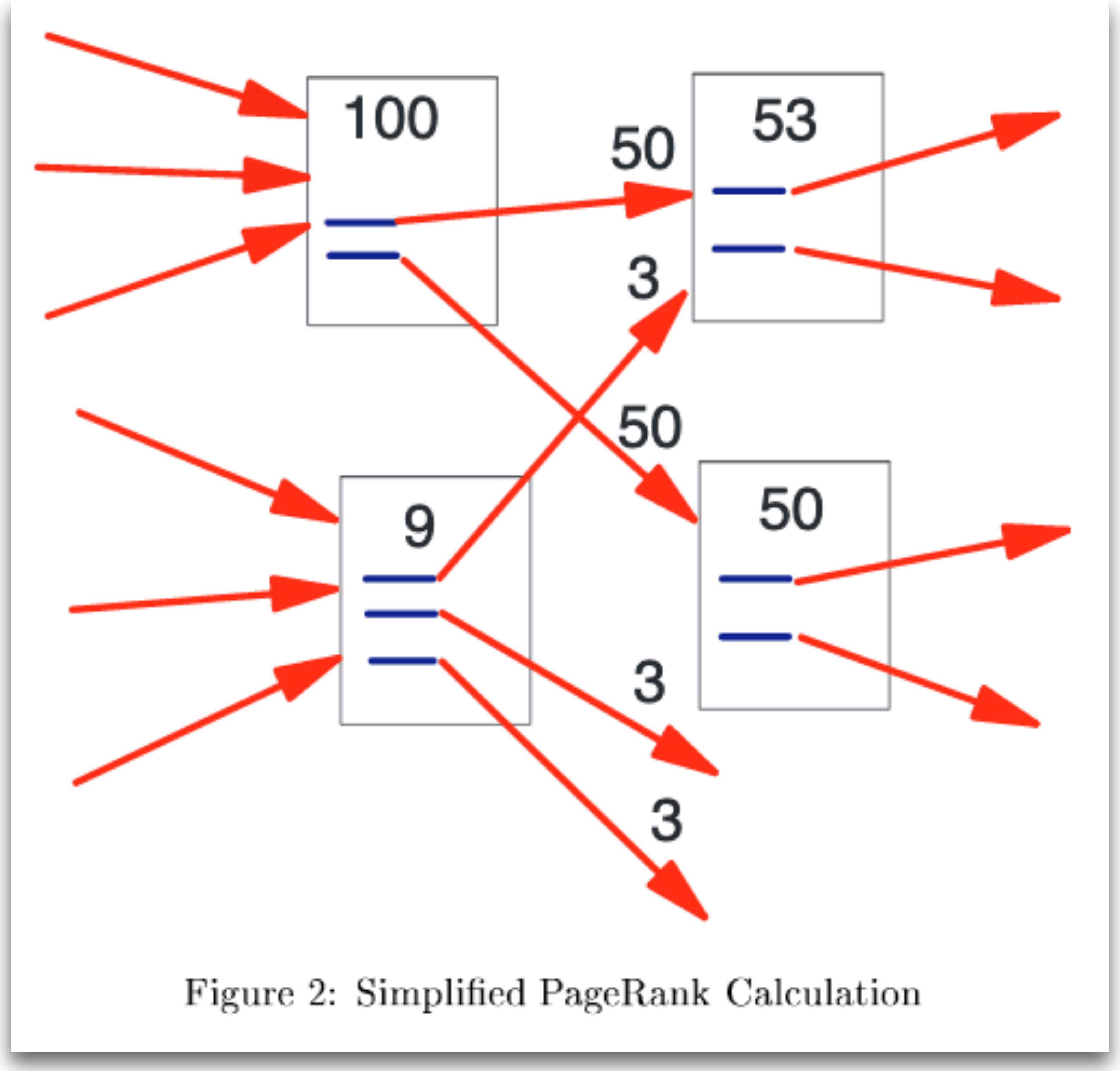


Figure 2: Simplified PageRank Calculation

HITS

- ➔ Query-dependent algorithm.
- ➔ Starts with the answer set (e.g. pages containing the keywords).
- ➔ Computes two scores for each page: authority and hub scores
  - ➔ Pages with many links pointing to them are called authorities.
  - ➔ Pages with many outgoing links are called hubs.

$$H(p) = \sum_{u \in S \;|\; p \rightarrow u} A(u) \; , \qquad A(p) = \sum_{v \in S \;|\; v \rightarrow p} H(v)$$

HITS Example

➔ From "Authoritative Sources in a Hyperlinked Environmen" (1999)

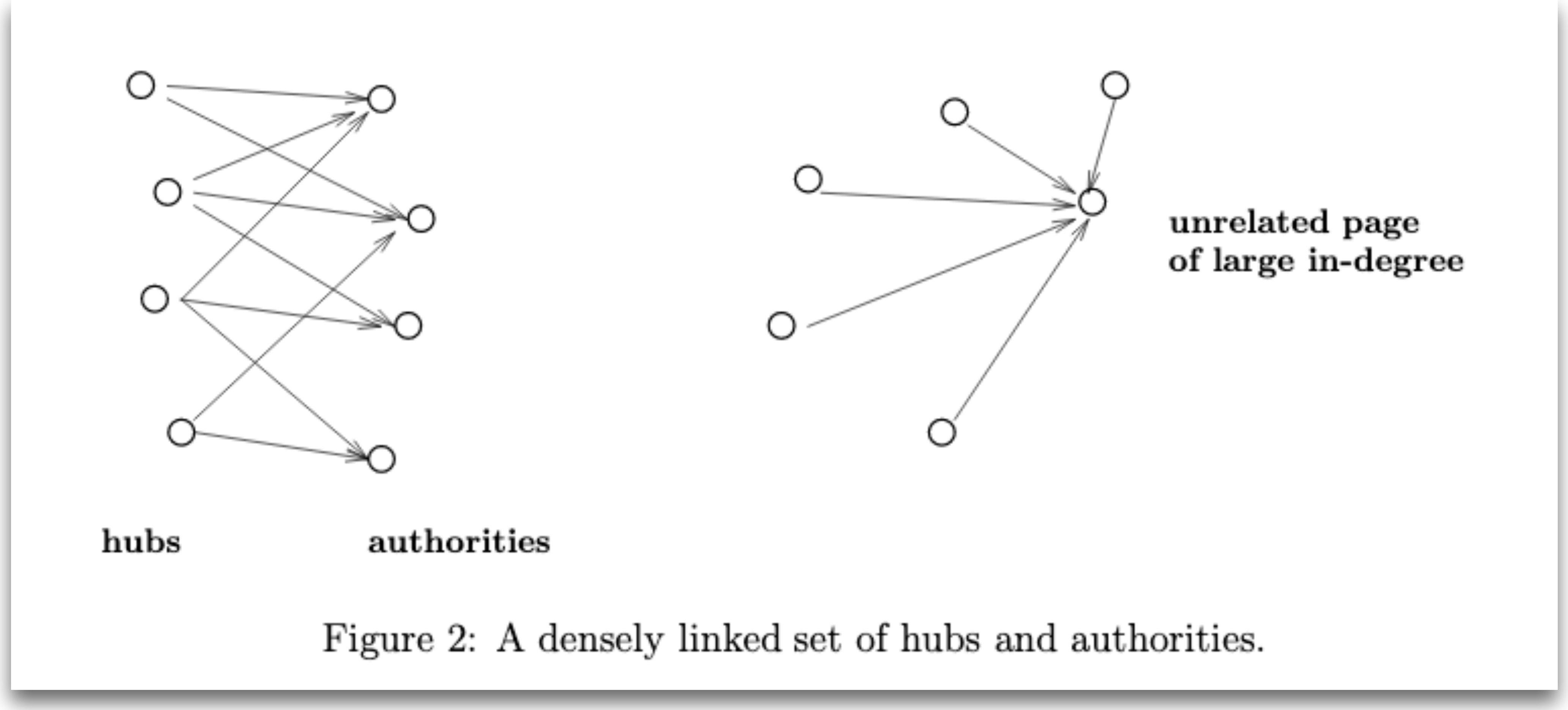


Figure 2: A densely linked set of hubs and authorities.

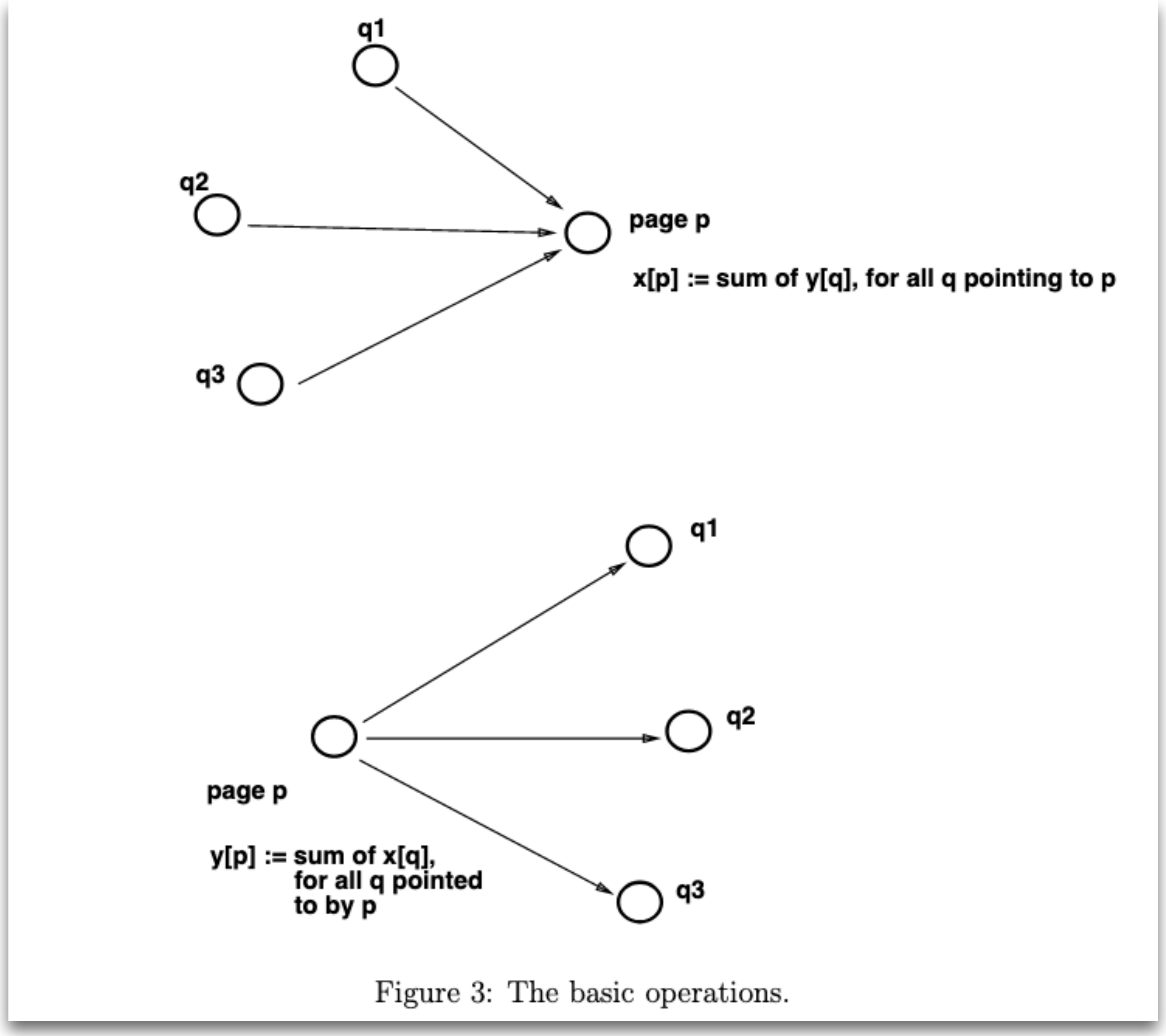


Figure 3: The basic operations.