Lesson 9

Hub, Authorities and Communities in Web Graph

.

Hub

- A hub is an index page that out-links to a number of content pages

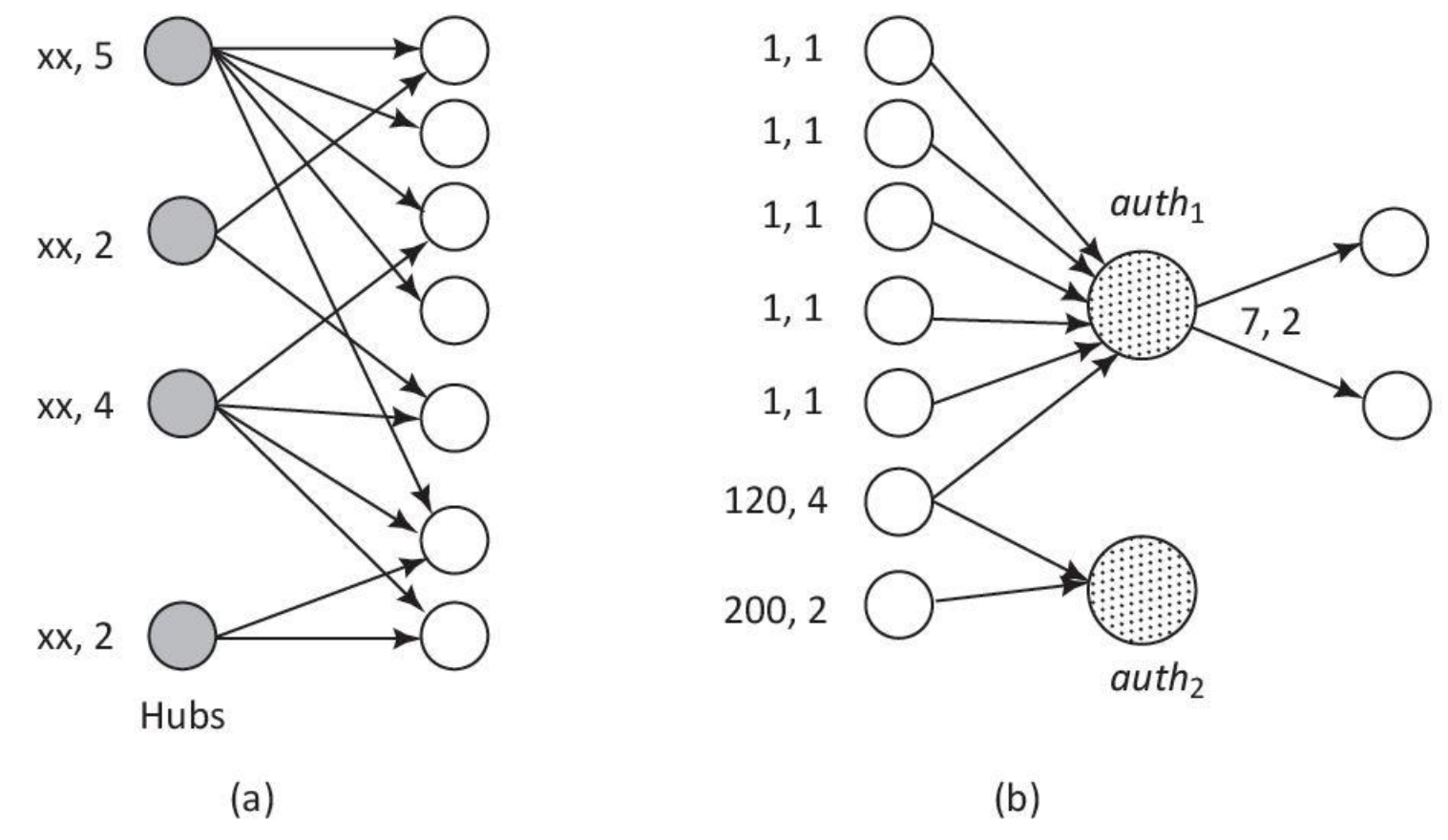
Authority

- Is a page that has recognition due to its useful, reliable and significant information.
- A content page is topic authority
- In-degrees (number of in-edges from other vertices) can be one of the measures for the authority. However, in-degrees do not distinguish between an in-link from a greater authority or

Authority based on In-degrees

- In-degrees do not distinguish between an in-link from a greater authority or lesser authority.

Figure 9.10 (a) Hubs (shaded circles) and (b) Authorities (dotted circles)



Authorities, *auth1* and *auth2*

- Figure 9.10(b)
- auth1 in-links from 6 vertices (in-degrees = 6) and
- auth2 has in-links to just 2 (in-degree = 2).

Authorities, *auth1* and auth2

- In-degrees not be a good measure
- The auth1 has link with six vertices with in-degrees = 1, 1, 1, 1, 1 and 120 (total = 125).
- Authority, auth2 has links with two vertices with in-degrees = 120 and 200 (total = 220). Auth2 has association with greater number of authorities.

Hypertext-Induced Topic Selection (HITS) algorithm

- Kleinberg (1998)
- Computes the hubs and authorities on a specific topic t.
- HITS analyses a sub-graph of web, which is relevant to t.

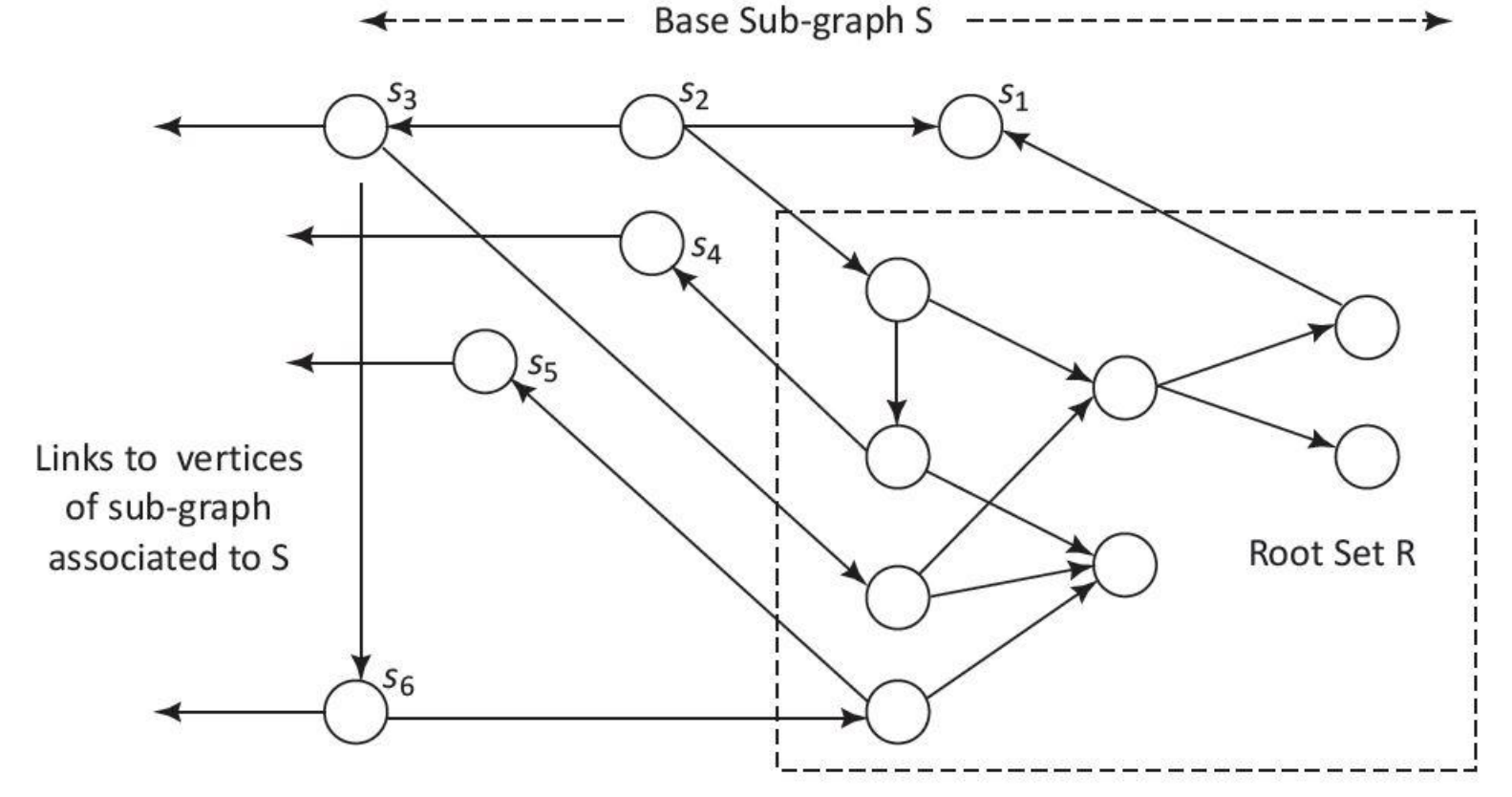
HITS Basis of Computations

- Hubs are the ones, which out-link to number of authorities, and
- Authorities are the ones, which in-link to number of hubs.
- A bipartite graph exists for the hubs and authorities

Spark GraphX

- Provides a set of fundamental operators such as subgraph, joinVertices and aggregateMessages
- Provides for computations using the property graphs
- Identifier in GraphX is 16-bit long unique-key. Edges have vertexIDs for corresponding source-destination paths.

Figure 9.11 Sub-graph for HITS consisting of root set R of pages and base sub-graph S including all the pages pointed to by any page of R.



Example 9.9

- Explains how are the hub and authority of pages in a given set of pages iterated and computed till the ranks do not change (within specified margin, that means until converges

Difference between HITS and PageRank

- HITS considers mutual reinforcement between authority and hub pages.
- PageRank ranks the pages just by authority and does not take into account distinctions between hubs and authorities
-

HITS and PageRank

- HITS considers the local neighbourhood between 4 to 8 pages surrounding the results of a query, whereas PageRank is applied to the entire web
- HITS depends on topic t, while PageRank is topic-independent.

Metrics for analyzing the Communities in Web Graph

- Web graph parameters, such as triangle count, clustering coefficient and K-neighbourhood

K-neighbourhood analysis for Analyzing Communities

- K-neighbourhood means the number of 1st neighbour nodes, 2nd neighbour nodes, and so on (K = 1, 2, 3, 4 and so on).

K-core analysis for analyzing Communities

- Means the number of cores within a marked area
- A core may consist of a triangle of connected vertices, a rectangle with interconnected edges and diagonals
- A core may also be a group of cores

Limitations of Link, Rank and Web Graph Analysis

1. Search engines rely on metatags or metadata of the documents, the rank enhances if metadata has biased information.
2. Search engines themselves may introduce bias while ranking the pages of clients rank changes.

Limitations of Link, Rank and Web Graph Analysis

3. A top authority may be a hub of pages on a different topic resulting in increased rank of the authority page
4. Topic drift and content evolution can affect the rank. Off-topic pages may return the authorities

- changes.

Limitations of Link, Rank and Web Graph Analysis

5. Mutually reinforcing affiliates or affiliated pages/sites can enhance each other's rank and authorities.
6. The ranks may be unstable as adding additional nodes may have greater influence in rank changes

Analytics using Spark GraphX

- Section 8.5
- Described functions for degree centralities, degree distribution, separation of degree, betweenness centralities, closeness centralities, neighbourhoods, strongly connected components, triangle counts, PageRank, shortest path

Analytics using Spark GraphX

- Breadth First Search (BFS), minimum spanning tree (forest), spectral clustering and cluster coefficient

Summary

We learnt:

- In-degrees and Out-degrees
- PageRank– The function measures the importance of each vertex in a graph
- PageRank iterative algorithm
- Hubs and Authorities; HITS algorithms
- Spark GrapgX for analytics of Web Graphs

End of Lesson 9 on

Hub, Authorities and Communities in Web Graph