Lesson 8

PageRank Analysis of a Web Structure

Web structure mining

- Process of discovering structure information from the web
- Based on the kind of structure-information present at the web resources

Hyperlinks

- Links exist between the web contents
- Link analysis :
- Is Page rank of a linked (web) higher or lower?
- Can the links be modeled as edges of graphs, structure of web as graph network, and applied the tools same as for graph analytics

Link Analytics

- Web graph analysis finding a link sending spam
- A set of links correspond to a hub
- Links corresponding to an authority
- A linked page has higher or lower authority compared to others

In-degree and Out-degree

- In-degree (visibility) of a link is the measure of number of in-links from other links
- Out-degree (luminosity) of a link is number of other links to which that link points

Page Rank (Authority)

- Each hyperlink in-links to a number of hyperlinks and out-links to a number of pages
- A page commanding higher authority (rank) has greater number of in-degrees than out-degrees

Page Authority

- One measure of a page authority can be in-degrees with respect to out-degrees.
- PageRank refers to the authority of the page measured in terms of number of times a link is sought after.

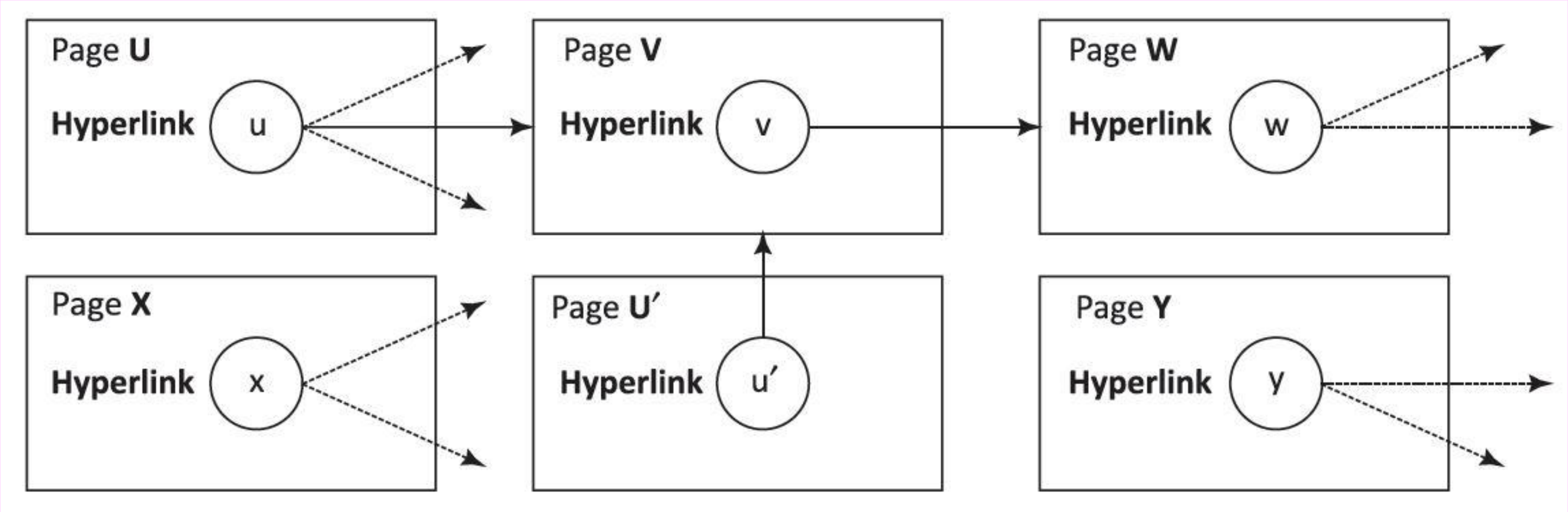
PageRank definition according to the new approach

- Page and co-authors (1998) defined a page ranking method
- Consider the entire web in place of local neighbourhood of the pages, and
- Consider the relative authority of the parent links (over children).

Pages U and U’ hyperlinks

- u and u’ out-linking to Page V
- Let Page U has three hyperlinks parenting three Pages, V one, W two, X two, U’ one, and Y two, respectively.
- Figure 9.8

Figure 9.8 Web structure with hyperlinks from a parent to one or more pages



Web Structure

- Let n = number of hyperlinks at the page U
- Assume u is a vector with elements u1, u2, … un.
- Each page Pg (u) has anchors, called hyperlinks.
- Page Pg (v) consists of text document with m number of hyperlinks.

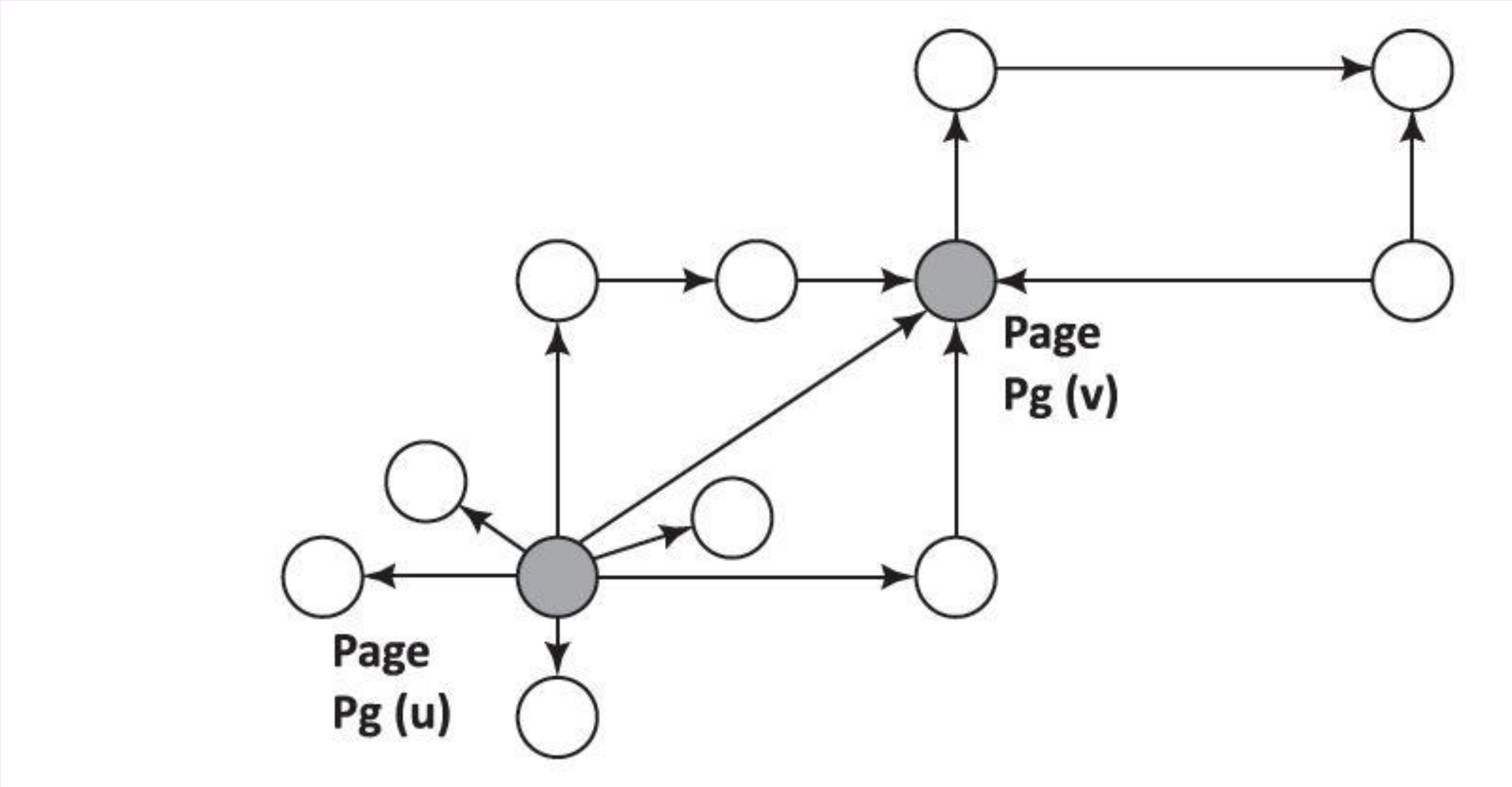
Out-Edges

- v is a vector with elements v1, v2, … vm
- The m is number of hyperlinks at Pg (v)
- A vertex u directs to another Page V
- A page Pg (v) may have number of hyperlinks directed by out-edges to other page Pg (w)

Authority

- Text at the hyperlink represents the property of a vertex u that describes the destination **V of the out-going edge.**
- A hyperlink in-between the pages represents the conferring of the authority.

Figure 9.9 Page Pg (v) in-links from Pg (u) and other pages



Web graph modeled as the web pages

- Page hyperlinks are the property of the graph node (vertex)
- Assume a Page, Pg (v) in-links from Pg (u), and Pg (u) out-linking similar to Pg (v), to total Nout [(Pg (u)] pages
- N_{out} for page U is 7 and for V is 1 in the figure. Number of in-linking N_{in} for page V is 4.

Computation of PageRank and PageRank Iteration

- Equation 9.21 initially suggested page rank, PR (based on in-degrees) of a page Pgv
- Rank computation algorithm then iterates and does the computations of rank-flowing (Example 9.7)

PageRank algorithm

- Using the relative authority of the parents over linked children
- Example 9.8 for rank computation algorithm iterating the rank flowing computations

PageRank Iteration using MapReduce

- Functions in Spark Grap*h*
- The method includes conversions to MapReduce functions and using HDFS compatible files. Functions PageRank (), ranksByUsername () do the computations using the PageRankObject.

GraphX consists of these functions

- GraphX Operators includes the functions (Section 8.5)
- Static PageRank algorithm runs for a fixed number of iterations, while dynamic

Convergence of Computations

- PageRank runs until the computed rank converges
- Convergence means that after certain iterations, the rank does not change significantly and any change remains within a pre-specified tolerance.

Topic Sensitive PageRank

- Topic-sensitive PageRank method uses surfing weights (probabilities) for the pages containing the topic or bag of words corresponding to a topic
- Compute the PageRank using the bias to rank R(v) and thus increase the effect of certain pages containing that topic or bag of words [Equation (9.29)]

Link Spam

- Effects of a link spam can be nullified using the topic-sensitive PageRank algorithm
- Link Spam tries to mislead the PageRank algorithm. A link spam attempts to make PageRank algorithm ineffective

Summary

We learnt:

- A page commanding higher authority (rank) has greater number of in-degrees than out-degrees
- Page algorithm computes PageRank using page iterations and does the computations of rank-flowing in the web links

End of Lesson 8 on

PageRank Analysis of a Web Structure