Course Project: MBA749 Fall 2023 **Due date: 10 Sep. 2023**

In this project we will use the concepts learnt in the course to analyze the internal Web structure of the news portal The Hindu. Document all challenges you faced while executing this project and their solution in the project report.

First you need to crawl the URL www.thehindu.com to obtain all URLs in the HTML of pages on the same subdomain. You may use the Web crawler Screaming Frog for this purpose. You need to gather all outlinks from the website of The Hindu, i.e., links on the site whose destination (target) is a different site. In Screaming Frog, you can get these links from Bulk Export → Links → All Outlinks. But if you are using a different crawler then it must have a similar option.

Open the edge list file and filter all "nofollow" links. This blog provides more information about the nofollow links and their role in search algorithms. In summary, nofollow links are not used to rank the sites therefore we will exclude them from analysis. We only need entries corresponding to the columns 'Source' and 'Destination' to construct the directed edge list. Import the filtered edge list in Gephi or any other network visualization software of your preference.

There may be multiple links between two pages, to resolve this you can select 'Sum' as Edges merge strategy. You can also check 'Self-loops' which allows you to construct edges to the page itself which is likely in the case of Web graph.

Select the largest connected component also known as giant component in the graph. Further, you may exclude the nodes with no outlinks and inlinks. For filtered graph run the following metrics: (a) In-degree (b) Out-degree (c) Betweenness centrality (d) Closeness centrality (e) PageRank, and (f) Modularity.

Modularity measures the extent to which like nodes are connected to like nodes in the network. It is used to cluster similar nodes in the network. For more details please refer to page 157-158 of the textbook. In Gephi you can compute it using Statistics → Community Detection → Modularity. The number of clusters should ideally be equal to the number of identifiable sections on the Website. You can generate a different number of clusters by tuning 'Resolution' parameter in the Modularity settings window. Mention the number of clusters you have selected based on modularity results.

Visualize your network by selecting a different color for each modularity class. Also, resize the nodes using PageRank. For the network layout select 'ForceAtlas 2'. You may have to change the default values in 'Tunining' section of ForceAtlas 2 to generate a good representation of the network. Also, experiment with different 'Min size' and 'Max size' values for a node until you see a network that is aesthetically appealing and the different modularity classes are well separated. A sample graph for your reference is reproduced below (see Figure 1). Include your network plot in the project report. Feel free to reposition nodes and modify the edge and node parameters for better visibility.

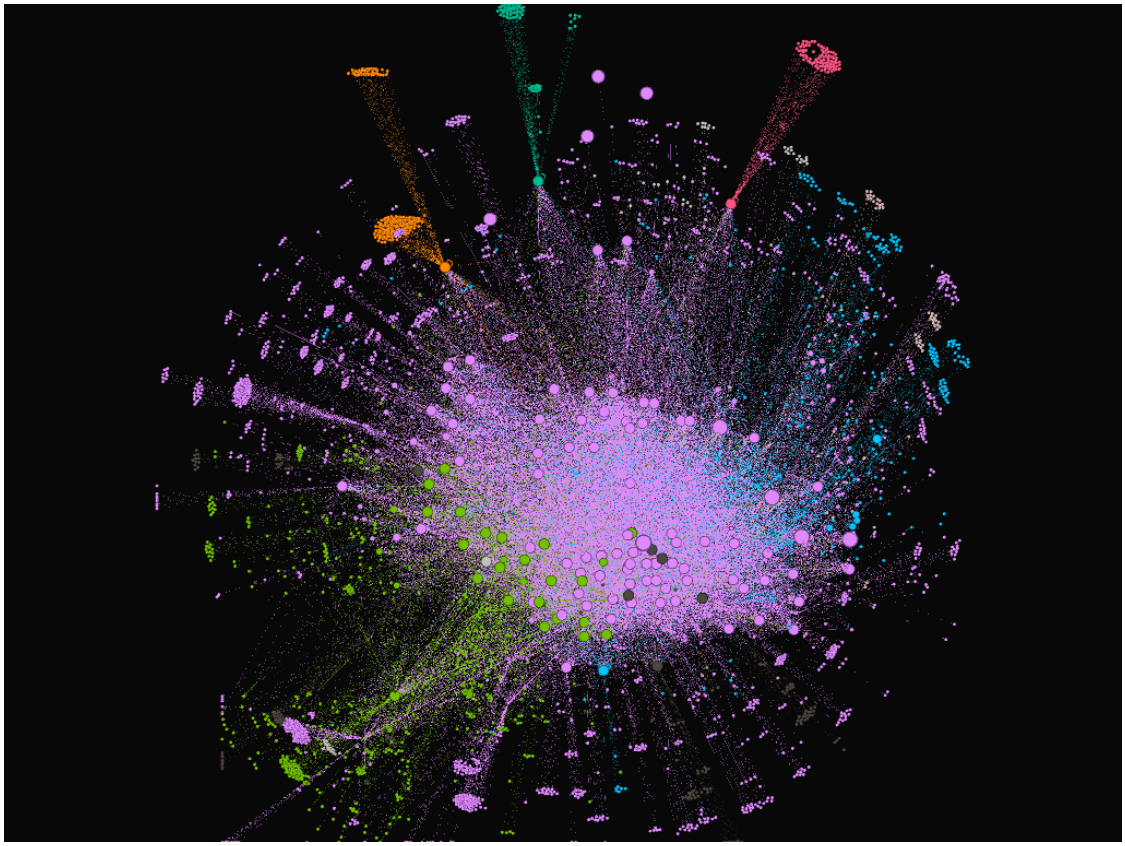

Figure 1. Different colors represent the different clusters present in the network

Now hide edges to visualize nodes with high PageRank and include this graph in your project report. A sample plot is given in Figure 2.

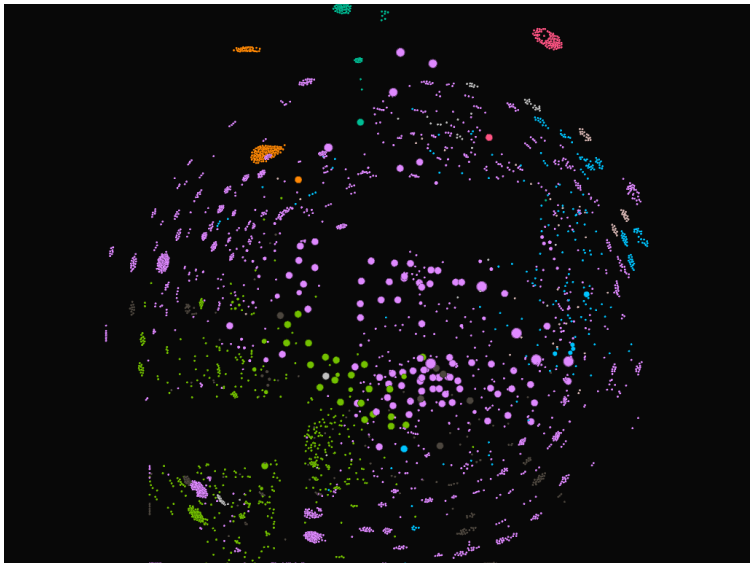

Figure 2. Network with edges where node size is proportion to PageRank

Identify the five largest nodes in the network and mention the hyperlinks corresponding to these nodes. While reporting these links exclude the sitewide links. A sitewide link is a link that appears on every page of the website. Examples of sitewide links are footer, sidebar, or site navigation menus. Also report top-5 nodes based on: In-degree, Out-degree, and Betweenness centrality. Mention how many nodes and edges are present in the largest connected component.

After completing all steps mentioned above, we have generated PageRank and other metrics internally for The Hindu news portal. Our universe of pages is limited to the sites (nodes) that appear in the network that we collected using crawler. Answer the following questions in your report.

Are there websites consistently ranked high based on all metrics?

Do the findings using network analysis match the empirical observation related to the web portal?

Is the Webpage with highest PageRank really the most prominent Webpage on the news portal? If not, what are the potential reasons for anomaly. Explain your answer.

Finally mention the contribution of each team member in the project and rate their performance on the scale of 1 to 10 where 1 is lowest and 10 is highest. Don’t forget to include the members name and roll no.