

MBA749A - Course Project

Gauri G Menon(200382) S Pradeep(200826)

Gephi Visualisation

Web Crawler (*crawled on 29-08-2023*)

The **Screaming Frog** web crawler was used to scrape the subdomain of “**www.thehindu.com**” for 500 sub-URLs (webpages). These produced **61,815** outlines. **2944** nofollow outlinks were filtered and removed from this file. The additional columns were removed and the Source and Destination columns were renamed into Source and Target columns to upload to Gephi as an edge list.

Graph Visualisation

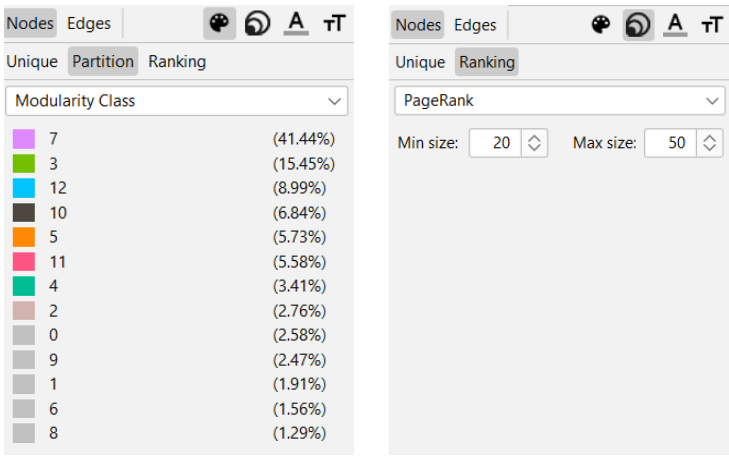
The filtered outlinks file was uploaded to Gephi as an edge list with a “**Sum**” Merge strategy for multi-edges. Self-loops were also allowed in the graph. This graph was then filtered to have only the giant component using the **Giant Component Topology** filter. On this network, in-degree, out-degree, PageRank, avg path length and modularity statistics were run on the network. This recorded the different centrality measures and the modularity of each node.

Modularity measures were run with resolution = 1.0 to lead to **13 clusters**, as 13 sections were observed in the subdomain of “The Hindu”.

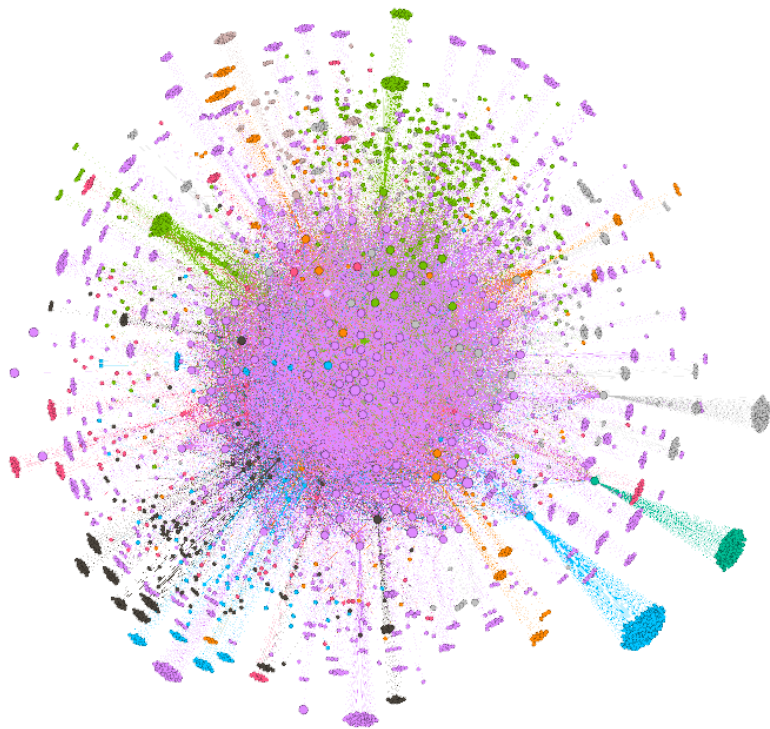
Clusters that are assumed to be part of the “The Hindu” subdomain are:

1. India
2. World
3. Opinion
4. Sports
5. Business
6. Sci-tech
7. Entertainment
8. Cities
9. States
10. Social Media/Contact
11. Advertisements
12. Life & Style
13. Food

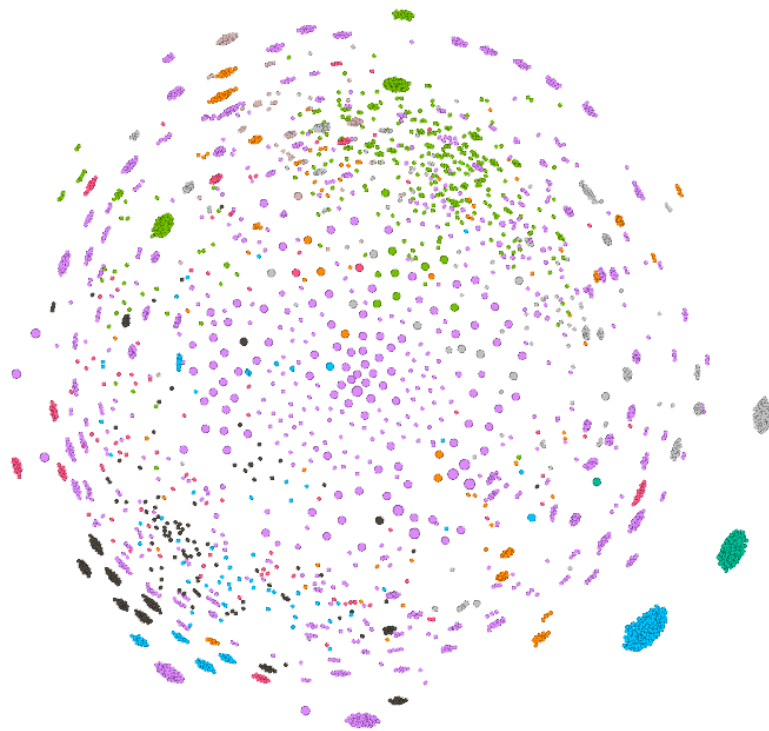
The graph was then visualised by fixing the node size in accordance with its **page rank**(min size = 20 and max size = 50) and its cluster colour according to its modularity. The graph was then visualised using the **Force Atlas 2** layout algorithm.



The visualised network looked as follows:

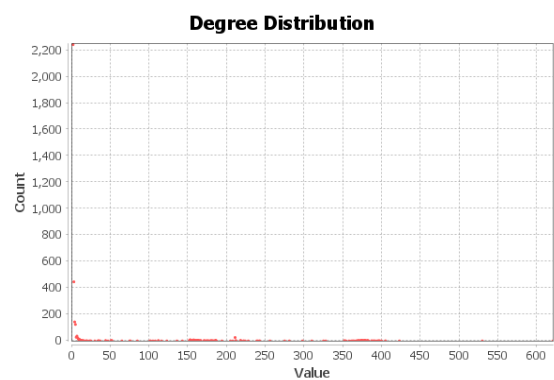
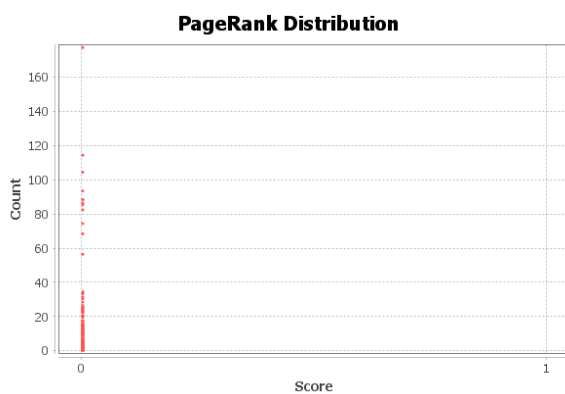


Network with nodes and edges

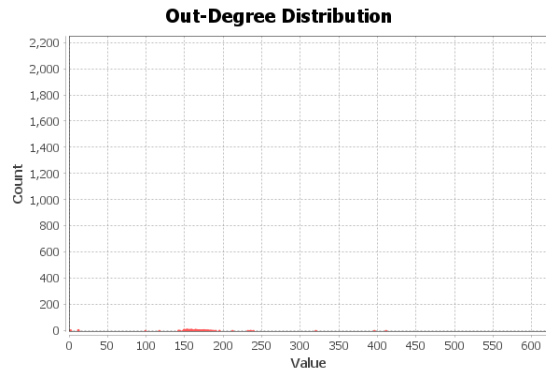
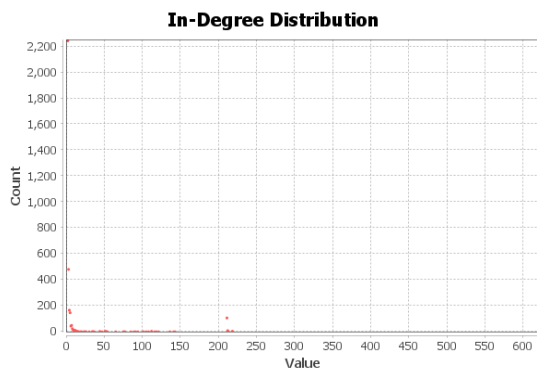


Network without edges and just nodes and node clusters

Network Report



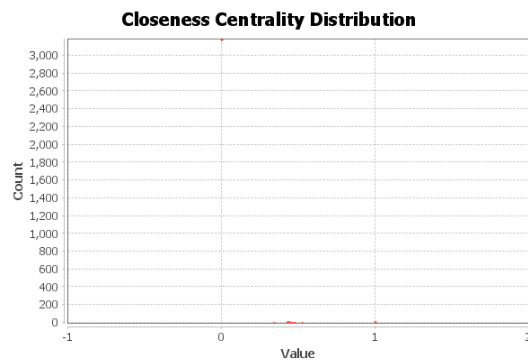
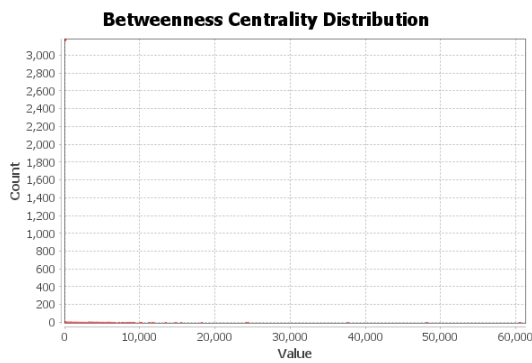
Average Degree: 10.381



Diameter: 4

Radius: 0

Average Path length: 2.303467608695955

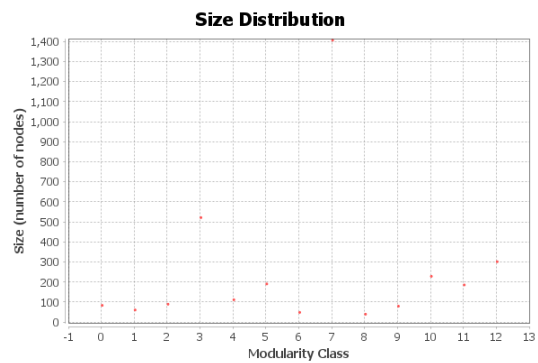


Resolution: 1.0

Modularity: 0.113

Modularity with resolution: 0.113

Number of Communities: 13



Network Analysis

Top-5 nodes

The node list with the centrality measures was then downloaded from Gephi. This node list was then arranged in descending order and manually labelled for sitewide links, which would boost its PageRank. It was seen that the **top 132** links were **sitewide links**, with some being **scripts**, **assets**(such as images and logos), and some **redirects**. These scripts (and assets), such as a Google fonts API, are used by all webpages to display the page contents on the browser. Some redirect links, such as <https://thehinduads.com/>, was not technically sitewide link but were redirected link from <http://www.thehinduclassifieds.in/>, which was a sitewide link. Some examples of such redirect links with high PageRank are: <https://thehinduads.com/>, <https://roofandfloor.thehindu.com/bangalore>, http://ad.apps.fm/Xqu5BIbECOREIqFMII5-hfE7og6fuV2o0Me0QdRqrE2LOAsYCovCD3BO4Tlr0voJxQxE023YRcpDAfPt4729MIV4EkBzhuiTQ6_P48Dm81I, <https://pay.hindu.com/esubspay/>

Excluding these redirect, script and sitewide links, the **top 5 nodes in page rank** are:

1. <https://frontline.thehindu.com/the-nation/education/controversy-should-teachers-share-their-political-opinions-with-students-sabyasachi-das-ashoka-university-karan-sangwan-unacademy/article67223267.ece>
2. <https://frontline.thehindu.com/the-nation/education/witch-hunt-against-tejaswini-de-sai-highlights-dangers-of-being-a-teacher-in-india-today/article67040371.ece>
3. <https://www.thehindu.com/news/international/china-releases-new-official-map-showing-territorial-claims/article67245869.ece>
4. <https://www.thehindu.com/sport/athletics/neeraj-chopra-wins-indias-first-gold-at-world-championships-in-javelin/article67242612.ece>
5. <https://sportstar.thehindu.com/athletics/world-athletics-championships-2023-medal-winners-full-list-india-results-neeraj-chopra-arshad-nadeem-parun-chaudhary-budapest/article67244325.ece>

The top 5 nodes in terms of **in-degree** are:

1. <https://frontline.thehindu.com/the-nation/education/witch-hunt-against-tejaswini-de-sai-highlights-dangers-of-being-a-teacher-in-india-today/article67040371.ece>
2. <https://frontline.thehindu.com/the-nation/education/controversy-should-teachers-share-their-political-opinions-with-students-sabyasachi-das-ashoka-university-karan-sangwan-unacademy/article67223267.ece>
3. <https://www.thehindu.com/news/international/china-releases-new-official-map-showing-territorial-claims/article67245869.ece>
4. <https://www.thehindu.com/sport/athletics/neeraj-chopra-wins-indias-first-gold-at-world-championships-in-javelin/article67242612.ece>

5. <https://sportstar.thehindu.com/athletics/world-athletics-championships-2023-medal-winners-full-list-india-results-neeraj-chopra-arshad-nadeem-parun-chaudhary-budapest/article67244325.ece>


The top 5 nodes in terms of **out-degree** are:

1. <https://www.thehindu.com/sci-tech/science/chandrayaan-3-isro-future-gaganyaan-rlv-sslv-sce-200/article67236348.ece>
2. <https://www.thehindu.com/news/national/isro-third-moon-mission-chandrayaan-3-package/article67079189.ece>
3. <https://www.thehindu.com/sci-tech/science/explained-why-did-chandrayaan-3-land-on-the-near-side-of-the-moon/article67235632.ece>
4. <https://www.thehindu.com/sci-tech/science/isro-shares-video-showing-pragyan-rover-roaming-around-shiv-shakti-point/article67238464.ece>
5. <https://www.thehindu.com/news/morning-digest-august-29-2023/article67245990.ece>

The top 5 nodes in terms of **betweenness centrality** are:

1. https://www.thehindu.com/topic/The_Hindu_Explains/
2. <https://www.thehindu.com/sci-tech/science/isro-shares-video-showing-pragyan-rover-roaming-around-shiv-shakti-point/article67238464.ece>
3. <https://www.thehindu.com/news/national/other-states/orissa-hc-orders-relocation-of-all-stray-dogs-from-campus-of-national-law-university/article67247179.ece>
4. <https://www.thehindu.com/news/national/andhra-pradesh/andhra-pradesh-congress-accuses-government-of-commercialising-medical-education-in-state-demands-repeal-of-gos-107-108/article67247473.ece>
5. <https://www.thehindu.com/news/morning-digest-august-23-2023/article67224785.ece>

The network finally contains **3405 nodes** and **35347 edges**. Link to the nodes table:

 MBA749A Project 1

Context ×	
Nodes:	3405
Edges:	35347
Directed Graph	

Are there websites consistently ranked high based on all metrics?

After converting the nodes table into a data frame, **each row was ranked** based on the different metrics(i.e. In-degree, out-degree, PageRank, closeness centrality, and betweenness centrality). These **ranks were added** and the nodes with the **lowest overall rank** were filtered to find websites that rank high in all metrics.

The top 5 websites according to these metrics are(excluding sitewide links):

- | | |
|--|-----------------|
| 1. https://www.thehindu.com/ | Rank Sum: 51.0 |
| 2. https://www.thehindu.com/opinion/ | Rank Sum: 74.5 |
| 3. https://www.thehindu.com/entertainment/ | Rank Sum: 125.5 |
| 4. https://www.thehindu.com/news/ | Rank Sum: 129.0 |
| 5. https://www.thehindu.com/sitemap/ | Rank Sum: 147.0 |

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

Some observations to compare the network analysis and empirical data:

- 1) Most **sitewide links**(excluding scripts, assets, and APIs) rank high in all metrics as they are the most **connected, central, and socially relevant**. In addition, it helps filter out prominent news in specific domains and is also accessible on all web pages.
- 2) Apart from accessible sitewide links, the most **prominent news** for the day(here; geo-political, sports, crime, and sci-tech) **ranks highest** in each metric.
- 3) Interestingly, "**sci-tech**" news seems to rank highest in the out-degree metric, and we can attribute that to science articles requiring the highest amount of references.
- 4) "**The Hindu Explains**" seems to have the highest betweenness centrality(a measure of how often a node acts as a bridge connecting other nodes in the network). This could be because it is one of the only nodes(not sitewide) that connects the home page to many premium and popular articles of all genres.

These observations prove that **network analysis largely matches empirical data**(except for inaccessible sitewide links, discussed in the next question). The analysis is consistent with observation based on the popularity, social relevance, and structure of the website.

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

The most **prominent** webpages, according to PageRank, are **scripts, assets, and APIs** that are used by “The Hindu” subdomain to host their websites and articles. These are definitely not the most prominent web pages on the news portal and, in fact, are not at all visible to the user. These are links which are used in the back end of each website.

	A	B	C	D	E	F	G	H
1	Id	indegree	outdegree	modularity_class	closnesscentrail	betweenesscent	pageranks	Type
2	https://static.cloudflareinsights.com/beacon.min.js/v8b253dfea2ab4077af8c6f58422dfbdf1689876627854	218	0	7	0	0	0.001136	Script
3	https://www.thehindu.com/cdn-cgi/scripts/7d0fa10a/cloudflare-static/rocket-loader.min.js	217	0	7	0	0	0.001134	Script
4	https://fonts.googleapis.com/css2?family=Merriweather+Sans:ital,wght@0,300,0,400,0,500,0,600,0,700,0	217	0	7	0	0	0.001129	Script
5	https://fonts.googleapis.com/css2?family=Playfair+Display:wght@400,600,700,800,900&display=swap	217	0	7	0	0	0.001129	Script
6	https://cdn.cxense.com/cx.coe.js	217	0	7	0	0	0.001129	Script
7	https://www.googletagmanager.com/ns.html?id=GTM-W5VV9N	217	0	7	0	0	0.001129	Script
8	https://thehinduads.com/	1	0	7	0	0	0.000936	Redirect
9	https://roofandfloor.thehindu.com/bangalore	1	0	7	0	0	0.000936	Redirect
10	http://ad.apps.fm/Xqu5BIBeCOREIqFMJ5-hfE7og6fuV2oOMeOQdRqrE2LOAsYCovCD3BO4Tlr0voJxQxI	1	0	7	0	0	0.000936	Redirect
11	https://pay.hindu.com/esubspa/	1	0	7	0	0	0.000936	Redirect
12	https://www.thehindu.com/subscription/	212	11	7	1	137.248512	0.00085	Sitewide
13	https://www.thehindu.com/	211	211	7	0.472384	61239.0897	0.000792	Sitewide
14	https://www.thehindu.com/news/	211	165	7	0.458019	60434.57221	0.000792	Sitewide
15	https://www.thehindu.com/opinion/	211	184	7	0.445842	11170.47704	0.000792	Sitewide
16	https://www.thehindu.com/life-and-style/	211	184	7	0.431651	7533.522805	0.000792	Sitewide
17	https://www.thehindu.com/entertainment/	211	183	7	0.43742	6586.554687	0.000792	Sitewide

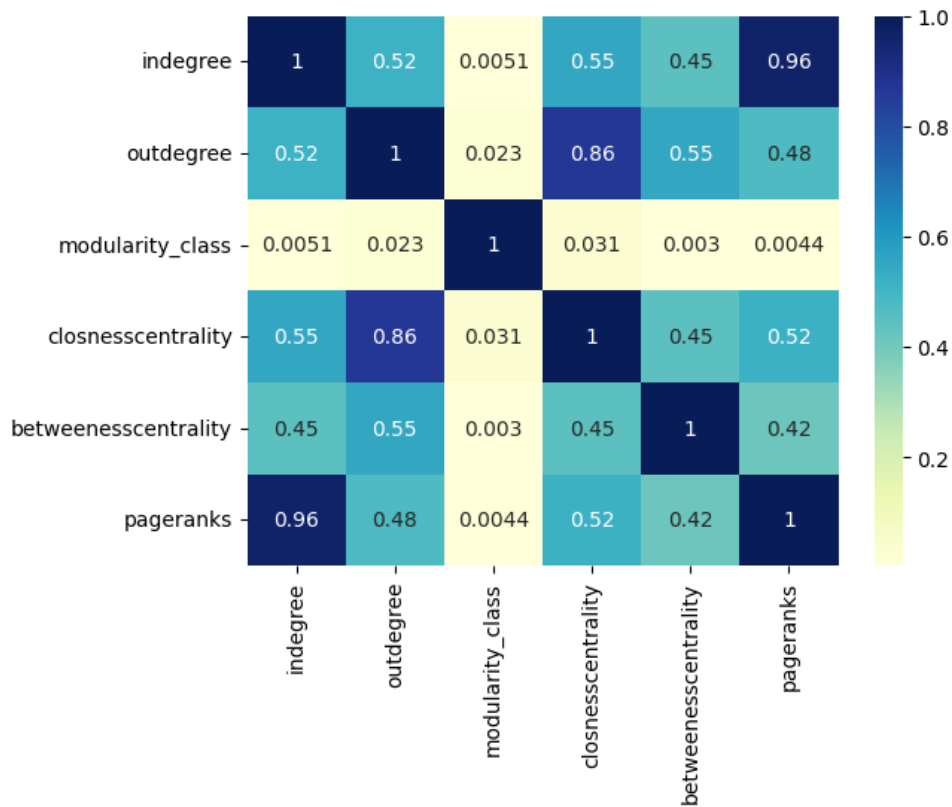
Additional Analysis MBA749A_Project1.ipynb

Centrality Feature Correlation

Some additional data analysis was performed on the node table data and the different centrality metrics to draw a more meaningful pattern. Firstly, a **correlation matrix** of the different centrality measures was plotted against each other.

We saw that the correlation value was very high for **PageRank and in-degree (96% correlation)**, indicating that both metrics are highly dependent. There was some correlation between **out-degree and closeness centrality (86% correlation)**, which could be attributed to more out-degree implying better connectivity to all nodes in the network.

The correlation matrix was as follows:

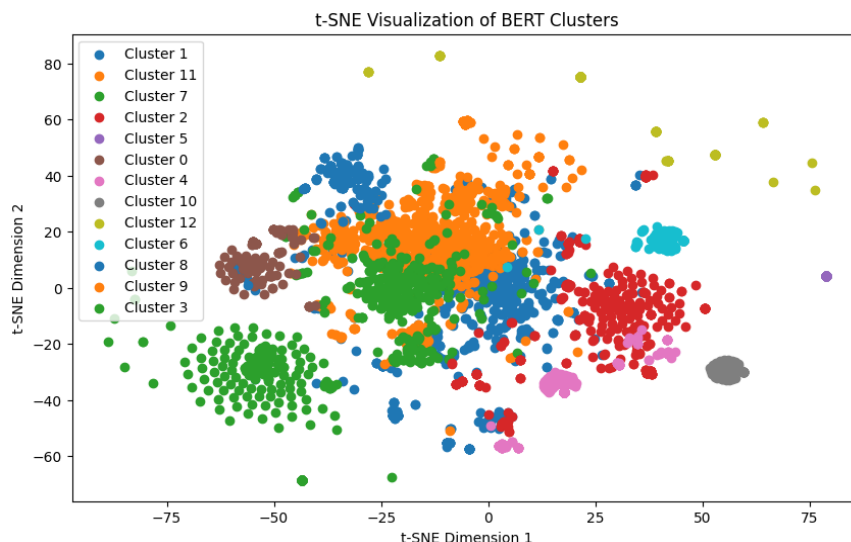


Semantic Analysis

The URLs were then split into **separate words** using the character delimiters. These URLs now became sentences. These sentences were **tokenized**, and filtered for only **meaningful words** from a “nlTK” dictionary. These were then **stemmed** and lemmatized.

These sentences were then converted into word embeddings using the **BERT transformer** and clustered based on **semantic similarity**. These clusters were **13** in number (matched with the same number of modularity classes).

These word embeddings were also then visualized using the **t-SNE algorithm** for dimensionality reduction, which visualizes it in 2D with a colour for each cluster.



Cluster Similarity

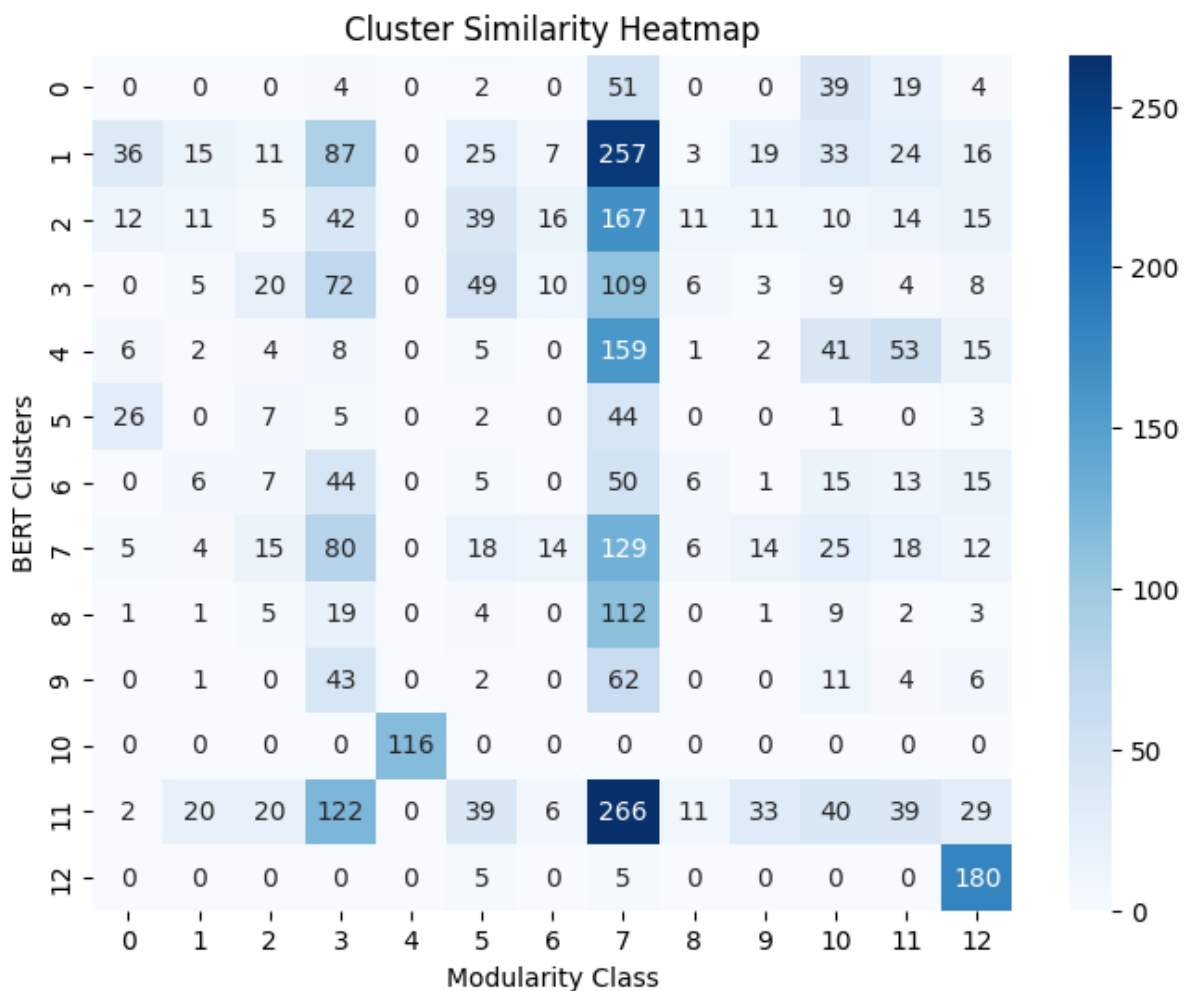
Cluster similarity was then calculated between the BERT clusters and the graph similarity communities(*this stems from an **assumption** that websites dealing with similar topics may also be linked together in the graph more closely and thus be more similar*).

We went on to calculate the **adjusted rand index(ARI)** for the 2 cluster sets created by Network analysis(Modularity Class) and Semantic Analysis(BERT). The ARI is a measure used to assess the similarity between two different clusterings or partitionings of data.

ARI = 1 indicates a perfect match between the clusterings.

Adjusted Rand Index (ARI): 0.052734048837947745

We have also made a **contingency heatmap** for the cluster similarity below:



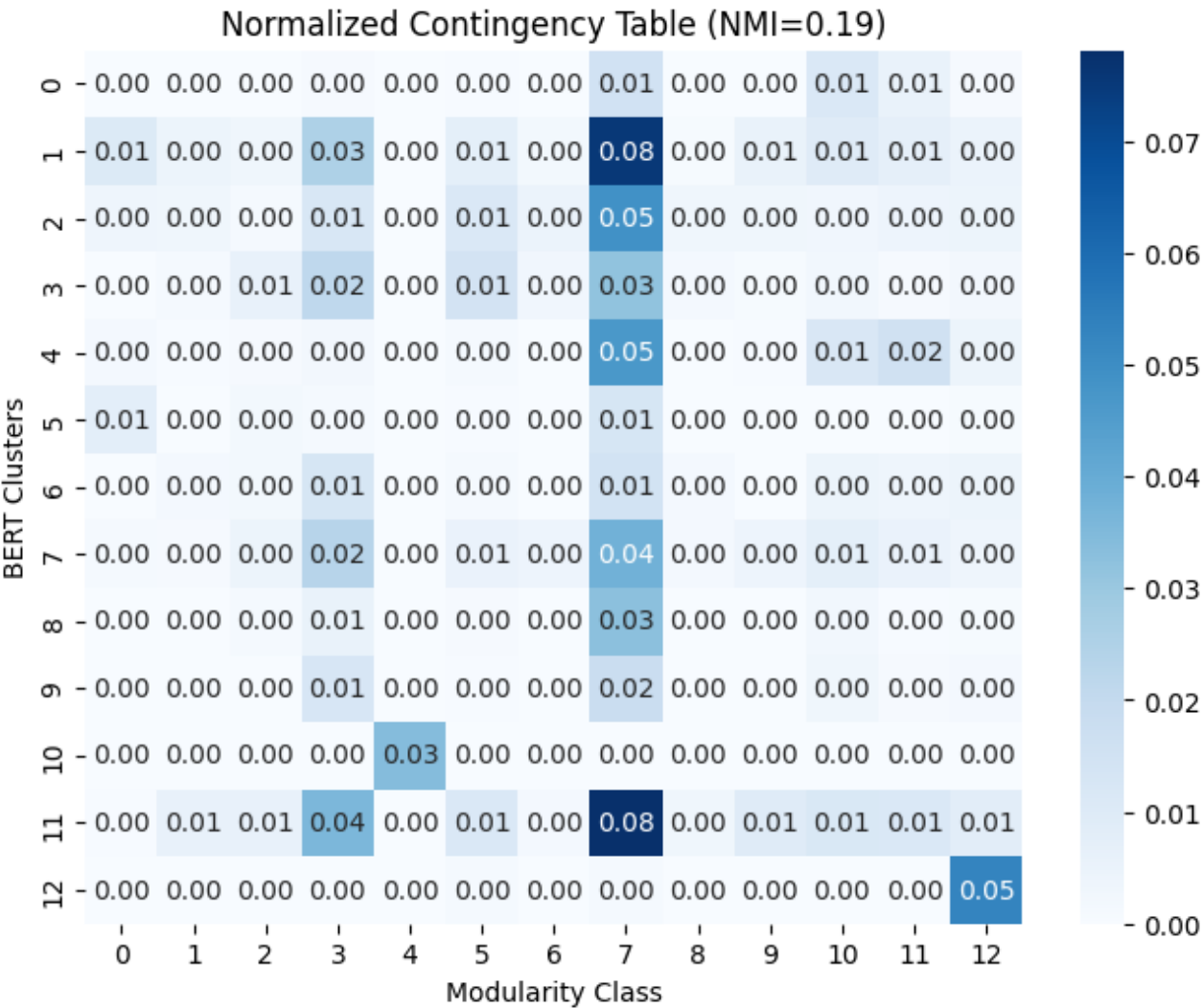
Here, it shows that cluster modularity cluster 7 matches highly with semantic analysis(BERT) clusters 1 and 11. Others show little similarity.

We recognised that this analysis might be skewed due to the relatively large size of modularity class 7. To counter this we calculated the normalised similarity metric of **Normalized Mutual Information (NMI)**. NMI normalizes the contingency table by considering the entropy of each cluster.

NMI = 1 indicates a perfect match between the clusterings.

NMI Score: 0.18815455533663553

We have also made a *contingency heatmap* for the normalised cluster similarity below:



Conclusion:

Except in a few instances, limited similarity is seen in Network and Semantic Analysis

Contribution

Team Member	Roll No	Contribution	Rating(/10)
Gauri G Menon	200382	Data Analysis, Data Plotting, Report Compilation	10
S Pradeep	200826	Gephi Visualisation, Empirical Analysis, Report Compilation	10