INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Social Media Analytics-I (MBA749A)

The Hindu News Portal Web Structure Analysis

Submitted by:

Sourit Saha
Roll No. 200998

Harsh Garg
Roll No. 200411

Abstract:

The digital landscape of news consumption has undergone significant transformations in recent years, with online news portals playing a pivotal role in disseminating information. This project delves into the intricate web structure of one such portal, The Hindu, employing a comprehensive analysis methodology. The primary aim is to uncover valuable insights into the portal's internal web connections and evaluate the prominence of webpages within this ecosystem.

The project initiates with web crawling, utilising tools such as Screaming Frog to harvest URLs from The Hindu's subdomain. A directed edge list is constructed by extracting outlinks and meticulously filtering out "nofollow" links. This list is the foundation for a network analysis involving metrics like In-degree, Out-degree, Betweenness centrality, Closeness centrality, and PageRank. Modularity analysis is employed to identify potential clusters within the web structure, mirroring the portal's diverse sections.

Visualising the network through Gephi and leveraging techniques such as ForceAtlas 2 layout, the project presents an aesthetically appealing representation where nodes are coloured by modularity classes and sized according to PageRank scores. The top nodes, identified based on multiple metrics, reveal significant webpages within The Hindu's architecture.

The project critically examines the relationship between web prominence, empirical observations, and network metrics. It evaluates whether websites consistently rank high based on these metrics and assesses the alignment between network analysis findings and real-world observations. Additionally, the project scrutinises whether the webpage with the highest PageRank reflects the most prominent content on the news portal, providing insights into potential anomalies.

This project offers a comprehensive analysis of The Hindu's web structure and underscores network analysis's utility in understanding news portals' digital architecture. It sheds light on the interconnectedness of web content within such platforms and the implications of web prominence metrics, contributing to a nuanced understanding of online news dissemination.

Table of Contents:

1. Introduction:

In an era characterised by the proliferation of digital information, the structure and organisation of websites play a pivotal role in how individuals access and consume online content. Understanding the intricate web structures of prominent online platforms is essential for comprehending the dissemination of information in the digital age. This project embarks on an in-depth exploration of the web structure of The Hindu, a well-established news portal, with the overarching goal of shedding light on its internal architecture and identifying key insights.

1.1 Objectives:

The primary objectives of this project are as follows:

- To crawl and analyse The Hindu's web domain to construct a comprehensive network of its internal web structure.
- To compute and evaluate various network metrics, including In-degree, Out-degree, Betweenness centrality, Closeness centrality, and PageRank, to assess the prominence and interconnectedness of webpages within the portal.
- To identify potential clusters or sections within The Hindu's website through modularity analysis, mirroring its diverse content offerings.
- To visually represent the web structure network, highlighting modularity classes and PageRank-based node sizes for enhanced understanding.
- To examine the alignment between web prominence metrics and empirical observations, particularly to the most prominent webpage on the news portal.

1.2 Significance:

The significance of this project lies in its multifaceted contribution to the fields of web analysis and information retrieval:

- Understanding Web Structure: This project offers insights into the complex web structure of a prominent news portal, which can be generalised to understand the organisation of other websites and online platforms.
- Web Prominence Evaluation: By assessing the web prominence of individual webpages using network metrics like PageRank, this project provides a framework for evaluating the importance of web content within a site.
- Modularity Analysis: Identifying potential clusters or sections within The Hindu's website enhances our understanding of its diverse content offerings, aiding content categorisation and user experience optimisation.
- Real-world Application: The findings of this project can inform web developers, content creators, and digital marketers about effective strategies for enhancing the visibility and accessibility of web content.
- Relevance to Digital Journalism: As online news portals continue to shape the media landscape; this project's insights contribute to a deeper understanding of how news is presented and accessed digitally.

This project ventures into the heart of The Hindu's digital ecosystem, unravelling its web structure, and aims to bridge the gap between empirical observations and network analysis metrics. It provides valuable insights for web analysts, content creators, and digital strategists while contributing to our broader understanding of information retrieval in the digital age.

2. Methodology:

The successful analysis of The Hindu news portal's web structure relied on a systematic methodology encompassing web crawling, data preprocessing, network analysis, and visualisation. This section elucidates the methods and tools employed in each project stage.

2.1 Web Crawling:

Web crawling is the foundational step in gathering data from The Hindu's web domain. To achieve this, the project utilised the versatile web crawling tool **Screaming Frog,** known for its efficiency in systematically extracting URLs and associated data from websites.

2.2 Data Preprocessing:

After obtaining the URLs from the web crawl, *data preprocessing* was done to ensure the data's quality and relevance. This stage involved the filtration of URLs to include only those contributing to the analysis.

2.3 Network Analysis:

With a refined dataset, the project proceeded to the core of the *network analysis*. This phase involved constructing a directed edge list and computing various network metrics to evaluate web prominence, connectivity, and clusters within The Hindu's web structure.

2.4 Visualisation:

Visualisation played a crucial role in enhancing understanding and communicating the findings. The network structure was visually represented using Gephi, incorporating techniques such as colour-coding, sizing nodes by PageRank, and employing the ForceAtlas 2 layout algorithm.

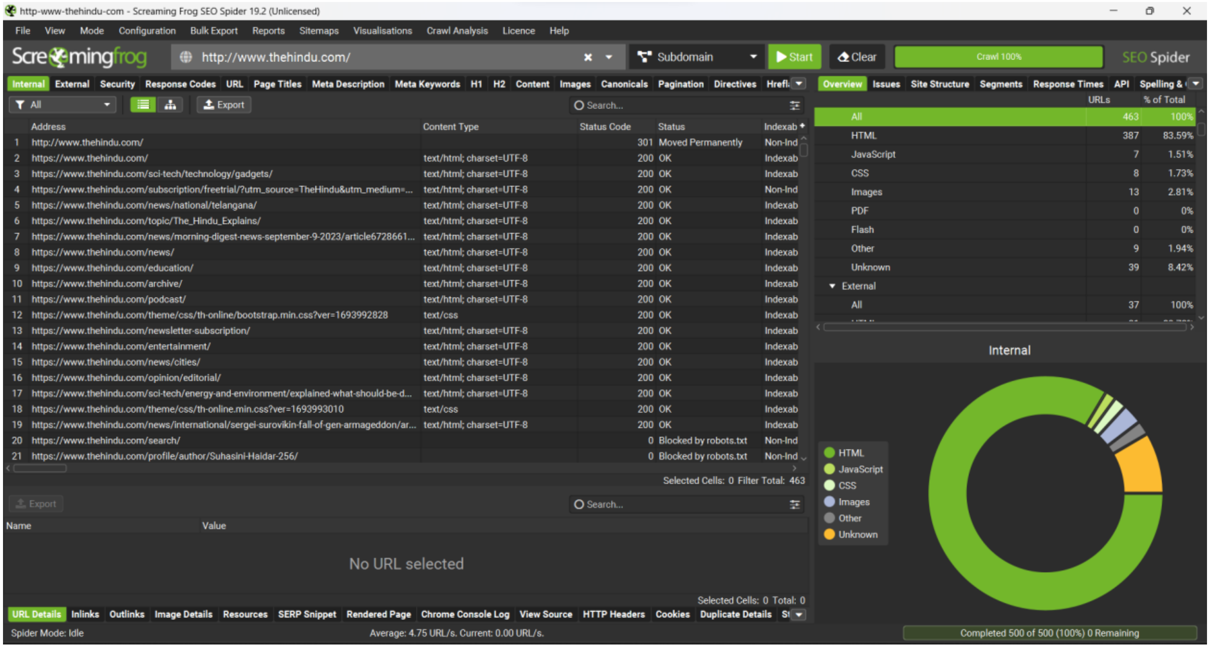
The methodology adopted for this project involved a systematic progression from web crawling to data preprocessing, network analysis, and, ultimately, visualisation. Each stage was characterised by its unique challenges and corresponding solutions. Screaming Frog, data manipulation tools, network analysis software, and visualisation techniques collectively facilitated the comprehensive exploration of The Hindu's web structure, yielding valuable insights into web prominence and content organisation.

3. Data Collection:

3.1 Crawling "The Hindu" Website:

Web crawling was initiated to systematically retrieve information from The Hindu's web domain. The project utilised the web crawling tool *Screaming Frog* for this purpose. The following steps were taken:

a. URL Entry: The starting point of the crawl was the main URL of The Hindu's website, www.thehindu.com.
b. Crawling Execution: Screaming Frog executed the crawl, systematically exploring the subdomain of The Hindu and collecting URLs, page titles, and other metadata.


http-www-thehindu-com - Screaming Frog SEO Spider 19.2 (Unlicensed)
File View Mode Configuration Bulk Export Reports Sitemaps Visualisations Crawl Analysis Licence Help
Screamingfrog
http://www.thehindu.com/
Subdomain
Start
Clear
Crawl 100%
SEO Spider
Internal External Security Response Codes URL Page Titles Meta Description Meta Keywords H1 H2 Content Images Canonicals Pagination Directives Hrefl
All
Export
Search...
Overview Issues Site Structure Segments Response Times API Spelling &
Internal
HTML
JavaScript
CSS
Images
Other
Unknown
Name
Value
No URL selected
URL Details Inlinks Outlinks Image Details Resources SERP Snippet Rendered Page Chrome Console Log View Source HTTP Headers Cookies Duplicate Details St

	Address	Content Type	Status Code	Status	Indexab
1	http://www.thehindu.com/		301	Moved Permanently	Non-Ind
2	https://www.thehindu.com/	text/html; charset=UTF-8	200	OK	Indexab
3	https://www.thehindu.com/sci-tech/technology/gadgets/	text/html; charset=UTF-8	200	OK	Indexab
4	https://www.thehindu.com/subscription/freetrial/?utm_source=TheHindu&utm_medium=...	text/html; charset=UTF-8	200	OK	Non-Ind
5	https://www.thehindu.com/news/national/telangana/	text/html; charset=UTF-8	200	OK	Indexab
6	https://www.thehindu.com/topic/The_Hindu_Explains/	text/html; charset=UTF-8	200	OK	Indexab
7	https://www.thehindu.com/news/morning-digest-news-september-9-2023/article6728661...	text/html; charset=UTF-8	200	OK	Indexab
8	https://www.thehindu.com/news/	text/html; charset=UTF-8	200	OK	Indexab
9	https://www.thehindu.com/education/	text/html; charset=UTF-8	200	OK	Indexab
10	https://www.thehindu.com/archive/	text/html; charset=UTF-8	200	OK	Indexab
11	https://www.thehindu.com/podcast/	text/html; charset=UTF-8	200	OK	Indexab
12	https://www.thehindu.com/theme/css/th-online/bootstrap.min.css?ver=1693992828	text/css	200	OK	Indexab
13	https://www.thehindu.com/newsletter-subscription/	text/html; charset=UTF-8	200	OK	Indexab
14	https://www.thehindu.com/entertainment/	text/html; charset=UTF-8	200	OK	Indexab
15	https://www.thehindu.com/news/cities/	text/html; charset=UTF-8	200	OK	Indexab
16	https://www.thehindu.com/opinion/editorial/	text/html; charset=UTF-8	200	OK	Indexab
17	https://www.thehindu.com/sci-tech/energy-and-environment/explained-what-should-be-d...	text/html; charset=UTF-8	200	OK	Indexab
18	https://www.thehindu.com/theme/css/th-online.min.css?ver=1693993010	text/css	200	OK	Indexab
19	https://www.thehindu.com/news/international/sergei-surovikin-fall-of-gen-armageddon/ar...	text/html; charset=UTF-8	200	OK	Indexab
20	https://www.thehindu.com/search/		0	Blocked by robots.txt	Non-Ind
21	https://www.thehindu.com/profile/author/Suhasini-Haidar-256/		0	Blocked by robots.txt	Non-Ind

Selected Cells: 0 Filter Total: 463

	URLs	% of Total
All	463	100%
HTML	387	83.59%
JavaScript	7	1.51%
CSS	8	1.73%
Images	13	2.81%
PDF	0	0%
Flash	0	0%
Other	9	1.94%
Unknown	39	8.42%
▾ External		
All	37	100%

Selected Cells: 0 Total: 0

Spider Mode: Idle

Average: 4.75 URL/s. Current: 0.00 URL/s.

Completed 500 of 500 (100%) 0 Remaining

3.2 Data Collection:

The data collected during the web crawling provided the foundation for subsequent analysis. The dataset included the following key details:

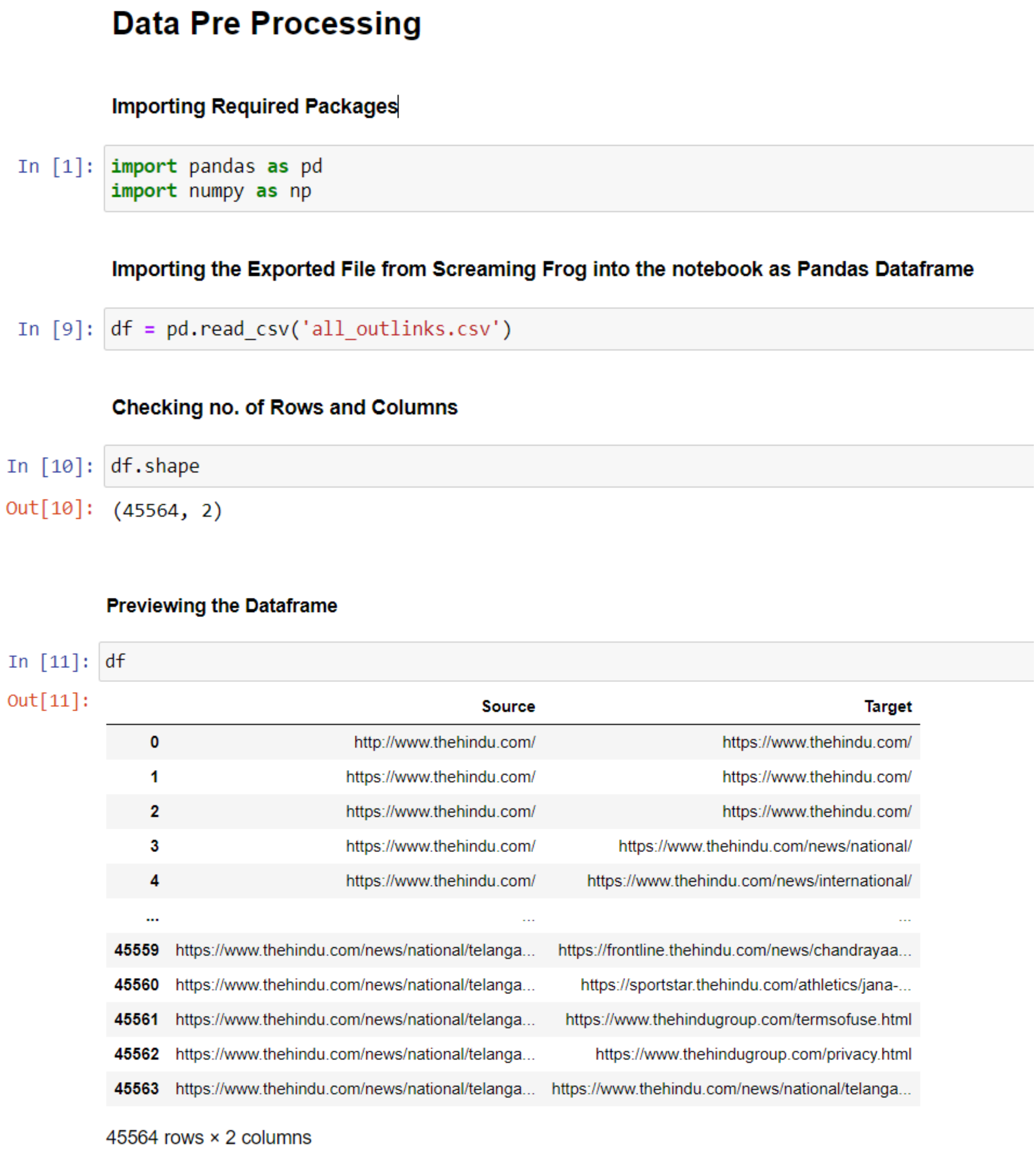
1. Number of URLs: The crawl yielded a substantial dataset of 500 URLs representing various web pages within The Hindu's subdomain.
2. Types of Links: The dataset encompassed various links, including internal links (within The Hindu's subdomain) and external links (pointing to other websites).
3. Filtered Data: As part of the data preprocessing stage, "nofollow" links were manually excluded from the dataset to ensure that the analysis focused exclusively on followed links contributing to web prominence.

Date of Web Crawling: 01st September 2023

The data collection process involved systematically crawling The Hindu's web domain using Screaming Frog. Despite the challenges posed by the website's vastness and the presence of "nofollow" links, the project successfully collected a comprehensive dataset. This dataset served as the raw material for subsequent stages of data preprocessing, network analysis, and visualisation, ultimately enabling a comprehensive exploration of The Hindu's web structure.

4. Data Preprocessing:

We have used rigorous Python programming to preprocess the data before feeding it into Gephi. Please follow the below code prompts:



Data Pre Processing

Importing Required Packages

```
In [1]: import pandas as pd
        import numpy as np
```

Importing the Exported File from Screaming Frog into the notebook as Pandas Dataframe

```
In [9]: df = pd.read_csv('all_outlinks.csv')
```

Checking no. of Rows and Columns

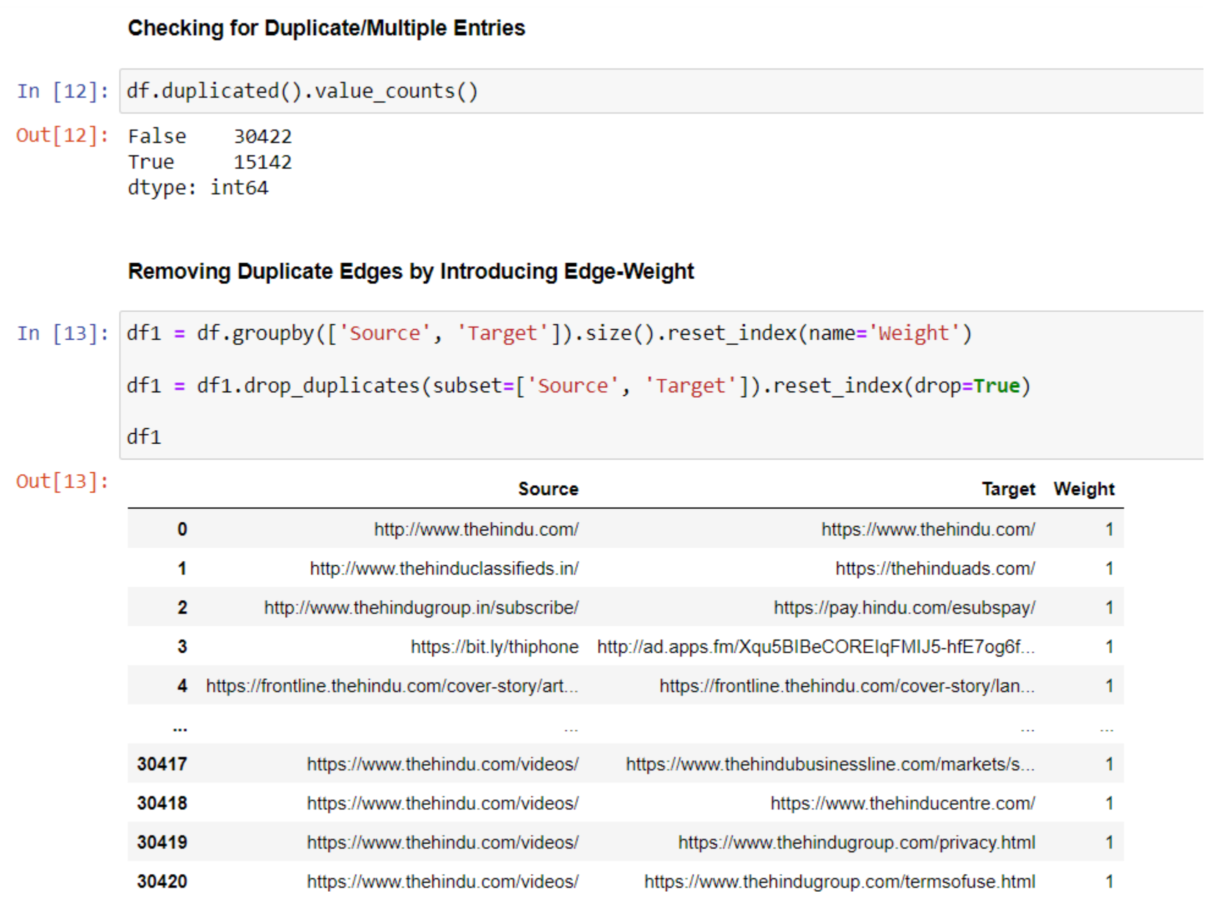
```
In [10]: df.shape
Out[10]: (45564, 2)
```

Previewing the Dataframe

```
In [11]: df
Out[11]:
```

	Source	Target
0	http://www.thehindu.com/	https://www.thehindu.com/
1	https://www.thehindu.com/	https://www.thehindu.com/
2	https://www.thehindu.com/	https://www.thehindu.com/
3	https://www.thehindu.com/	https://www.thehindu.com/news/national/
4	https://www.thehindu.com/	https://www.thehindu.com/news/international/
...	...	...
45559	https://www.thehindu.com/news/national/telanga...	https://frontline.thehindu.com/news/chandrayaa...
45560	https://www.thehindu.com/news/national/telanga...	https://sportstar.thehindu.com/athletics/jana-...
45561	https://www.thehindu.com/news/national/telanga...	https://www.thehindugroup.com/termsofuse.html
45562	https://www.thehindu.com/news/national/telanga...	https://www.thehindugroup.com/privacy.html
45563	https://www.thehindu.com/news/national/telanga...	https://www.thehindu.com/news/national/telanga...

45564 rows × 2 columns



Checking for Duplicate/Multiple Entries

```
In [12]: df.duplicated().value_counts()
```

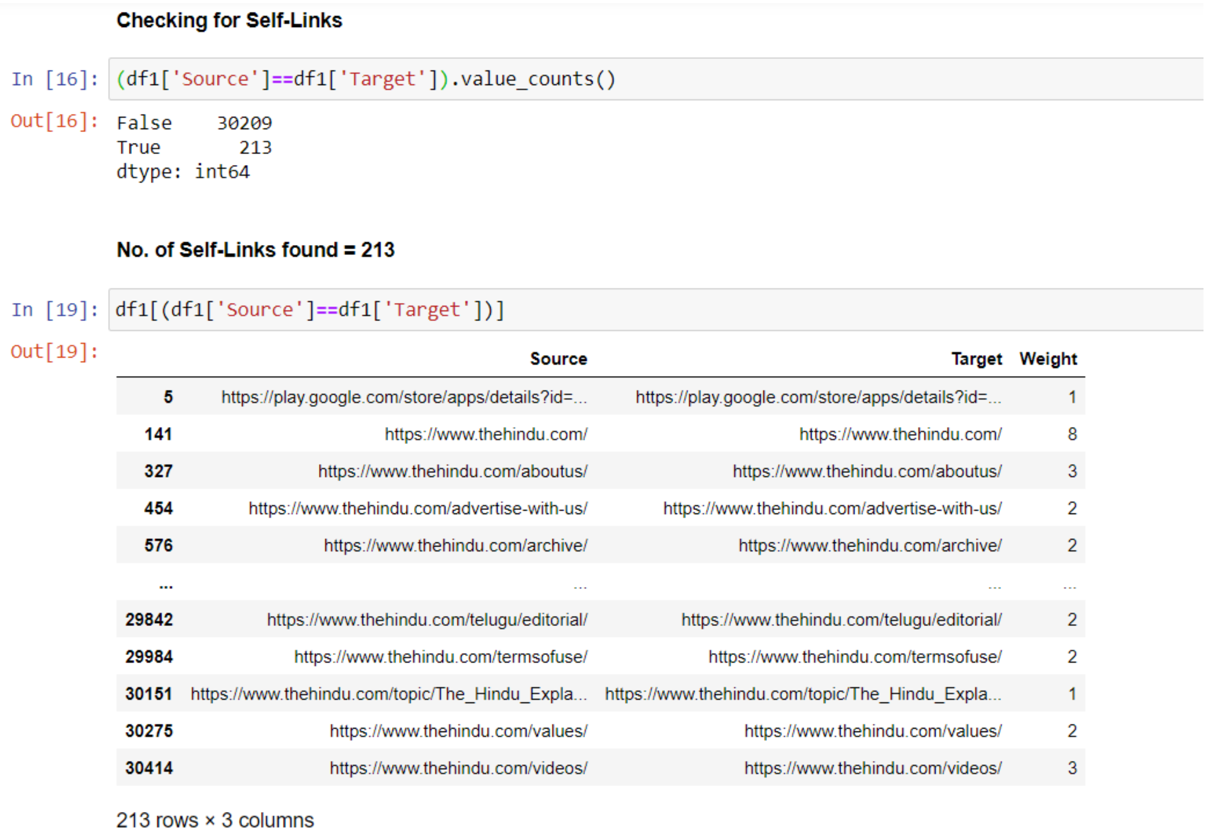
```
Out[12]: False    30422
         True     15142
         dtype: int64
```

Removing Duplicate Edges by Introducing Edge-Weight

```
In [13]: df1 = df.groupby(['Source', 'Target']).size().reset_index(name='Weight')

         df1 = df1.drop_duplicates(subset=['Source', 'Target']).reset_index(drop=True)

         df1
```

Out[13]:

	Source	Target	Weight
0	http://www.thehindu.com/	https://www.thehindu.com/	1
1	http://www.thehinduclassifieds.in/	https://thehinduads.com/	1
2	http://www.thehindugroup.in/subscribe/	https://pay.hindu.com/esubspay/	1
3	https://bit.ly/thiphone	http://ad.apps.fm/Xqu5BIBeCOREIqFMIJ5-hfE7og6f...	1
4	https://frontline.thehindu.com/cover-story/art...	https://frontline.thehindu.com/cover-story/lan...	1
...	...	...	...
30417	https://www.thehindu.com/videos/	https://www.thehindubusinessline.com/markets/s...	1
30418	https://www.thehindu.com/videos/	https://www.thehinducentre.com/	1
30419	https://www.thehindu.com/videos/	https://www.thehindugroup.com/privacy.html	1
30420	https://www.thehindu.com/videos/	https://www.thehindugroup.com/termsofuse.html	1



Checking for Self-Links

```
In [16]: (df1['Source']==df1['Target']).value_counts()
```

```
Out[16]: False    30209
         True       213
         dtype: int64
```

No. of Self-Links found = 213

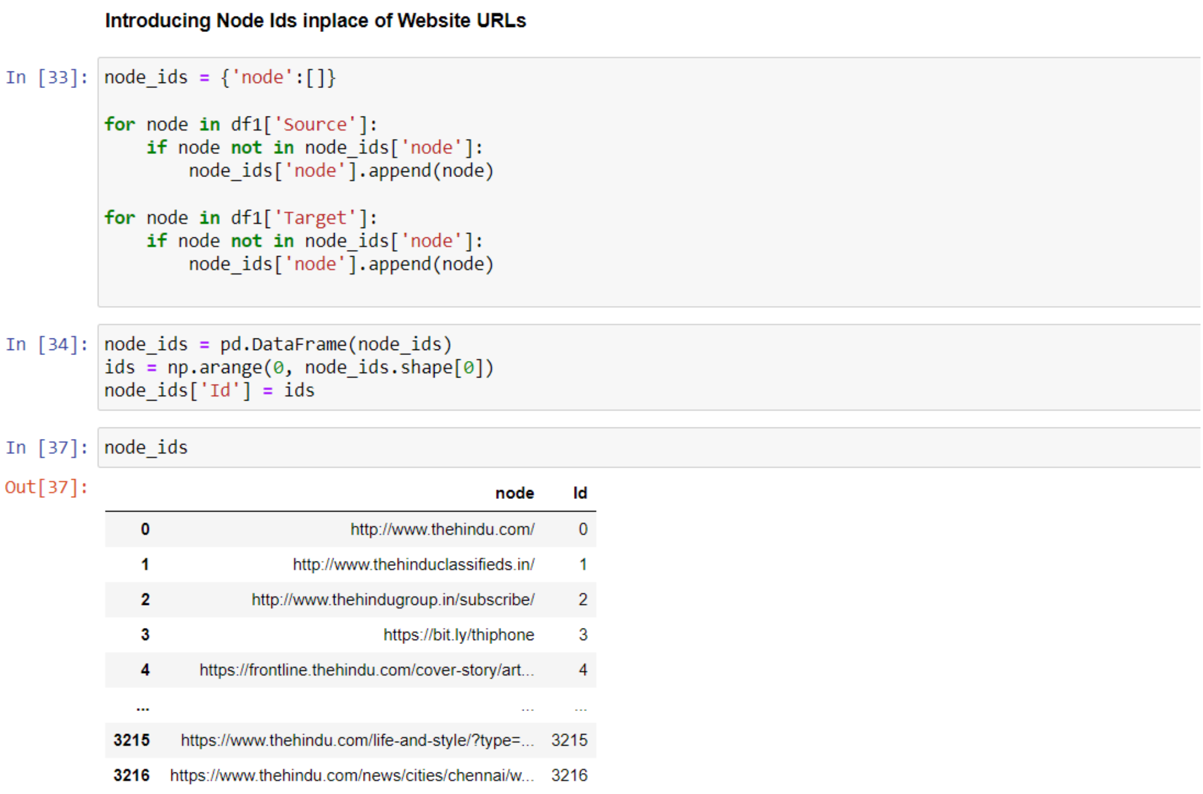
```
In [19]: df1[(df1['Source']==df1['Target'])]
```

Out[19]:

	Source	Target	Weight
5	https://play.google.com/store/apps/details?id=...	https://play.google.com/store/apps/details?id=...	1
141	https://www.thehindu.com/	https://www.thehindu.com/	8
327	https://www.thehindu.com/aboutus/	https://www.thehindu.com/aboutus/	3
454	https://www.thehindu.com/advertise-with-us/	https://www.thehindu.com/advertise-with-us/	2
576	https://www.thehindu.com/archive/	https://www.thehindu.com/archive/	2
...	...	...	...
29842	https://www.thehindu.com/telugu/editorial/	https://www.thehindu.com/telugu/editorial/	2
29984	https://www.thehindu.com/termsofuse/	https://www.thehindu.com/termsofuse/	2
30151	https://www.thehindu.com/topic/The_Hindu_Expla...	https://www.thehindu.com/topic/The_Hindu_Expla...	1
30275	https://www.thehindu.com/values/	https://www.thehindu.com/values/	2
30414	https://www.thehindu.com/videos/	https://www.thehindu.com/videos/	3

213 rows × 3 columns

We checked for islands in the Network but couldn't find any. This was later confirmed when we got only 1 Connected Component in the Network in Gephi.



Introducing Node Ids inplace of Website URLs

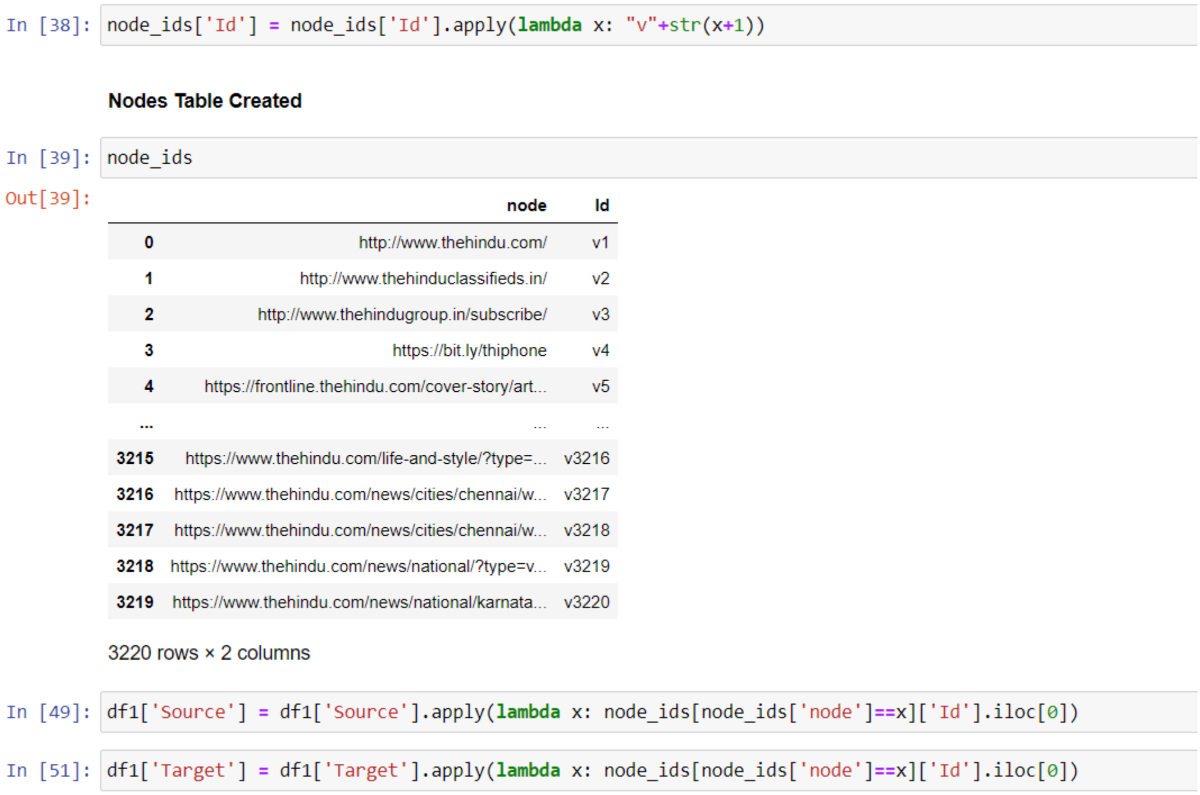
In [33]:
```
node_ids = {'node':[]}

for node in df1['Source']:
    if node not in node_ids['node']:
        node_ids['node'].append(node)

for node in df1['Target']:
    if node not in node_ids['node']:
        node_ids['node'].append(node)
```

In [34]:
```
node_ids = pd.DataFrame(node_ids)
ids = np.arange(0, node_ids.shape[0])
node_ids['Id'] = ids
```

In [37]:
```
node_ids
```

Out[37]:

	node	Id
0	http://www.thehindu.com/	0
1	http://www.thehinduclassifieds.in/	1
2	http://www.thehindugroup.in/subscribe/	2
3	https://bit.ly/thiphone	3
4	https://frontline.thehindu.com/cover-story/art...	4
...	...	...
3215	https://www.thehindu.com/life-and-style/?type=...	3215
3216	https://www.thehindu.com/news/cities/chennai/w...	3216



In [38]:
```
node_ids['Id'] = node_ids['Id'].apply(lambda x: "v"+str(x+1))
```

Nodes Table Created

In [39]:
```
node_ids
```

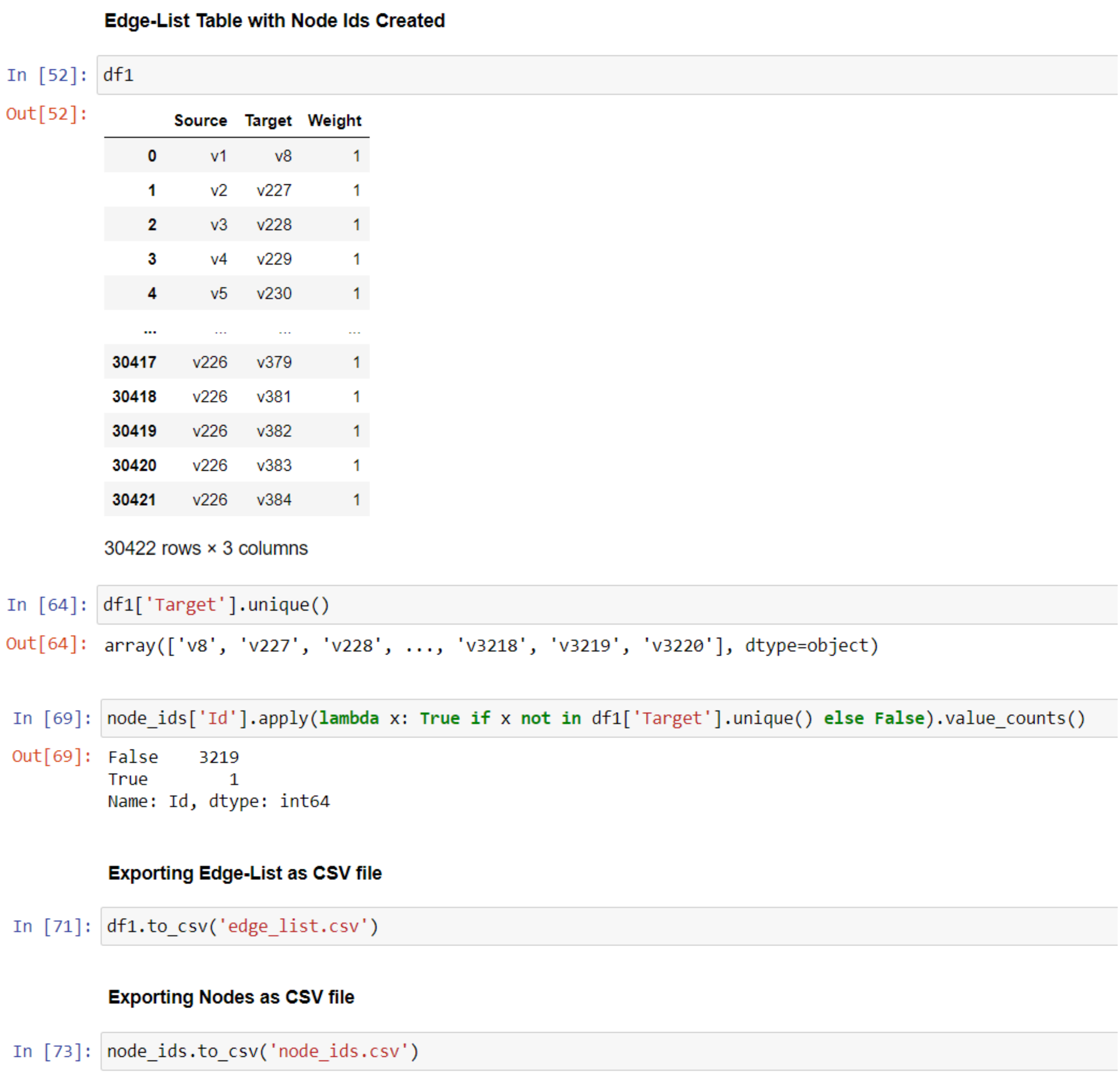
Out[39]:

	node	Id
0	http://www.thehindu.com/	v1
1	http://www.thehinduclassifieds.in/	v2
2	http://www.thehindugroup.in/subscribe/	v3
3	https://bit.ly/thiphone	v4
4	https://frontline.thehindu.com/cover-story/art...	v5
...	...	...
3215	https://www.thehindu.com/life-and-style/?type=...	v3216
3216	https://www.thehindu.com/news/cities/chennai/w...	v3217
3217	https://www.thehindu.com/news/cities/chennai/w...	v3218
3218	https://www.thehindu.com/news/national/?type=v...	v3219
3219	https://www.thehindu.com/news/national/karnata...	v3220

3220 rows × 2 columns

In [49]:
```
df1['Source'] = df1['Source'].apply(lambda x: node_ids[node_ids['node']==x]['Id'].iloc[0])
```

In [51]:
```
df1['Target'] = df1['Target'].apply(lambda x: node_ids[node_ids['node']==x]['Id'].iloc[0])
```



Edge-List Table with Node Ids Created

```
In [52]: df1
```

Out[52]:

	Source	Target	Weight
0	v1	v8	1
1	v2	v227	1
2	v3	v228	1
3	v4	v229	1
4	v5	v230	1
...	...	...	...
30417	v226	v379	1
30418	v226	v381	1
30419	v226	v382	1
30420	v226	v383	1
30421	v226	v384	1

30422 rows × 3 columns

```
In [64]: df1['Target'].unique()

Out[64]: array(['v8', 'v227', 'v228', ..., 'v3218', 'v3219', 'v3220'], dtype=object)
```

```
In [69]: node_ids['Id'].apply(lambda x: True if x not in df1['Target'].unique() else False).value_counts()

Out[69]: False    3219
         True        1
         Name: Id, dtype: int64
```

Exporting Edge-List as CSV file

```
In [71]: df1.to_csv('edge_list.csv')
```

Exporting Nodes as CSV file

```
In [73]: node_ids.to_csv('node_ids.csv')
```

Thus, we preprocessed the data as much as possible and fed it into Gephi for further analysis.

Please find the Code File here for reference:

https://drive.google.com/file/d/1Jp9JbfEHnORFmFw52V-Gyn_WjIX0LTD-/view?usp=drive_link

5. Data Analysis:

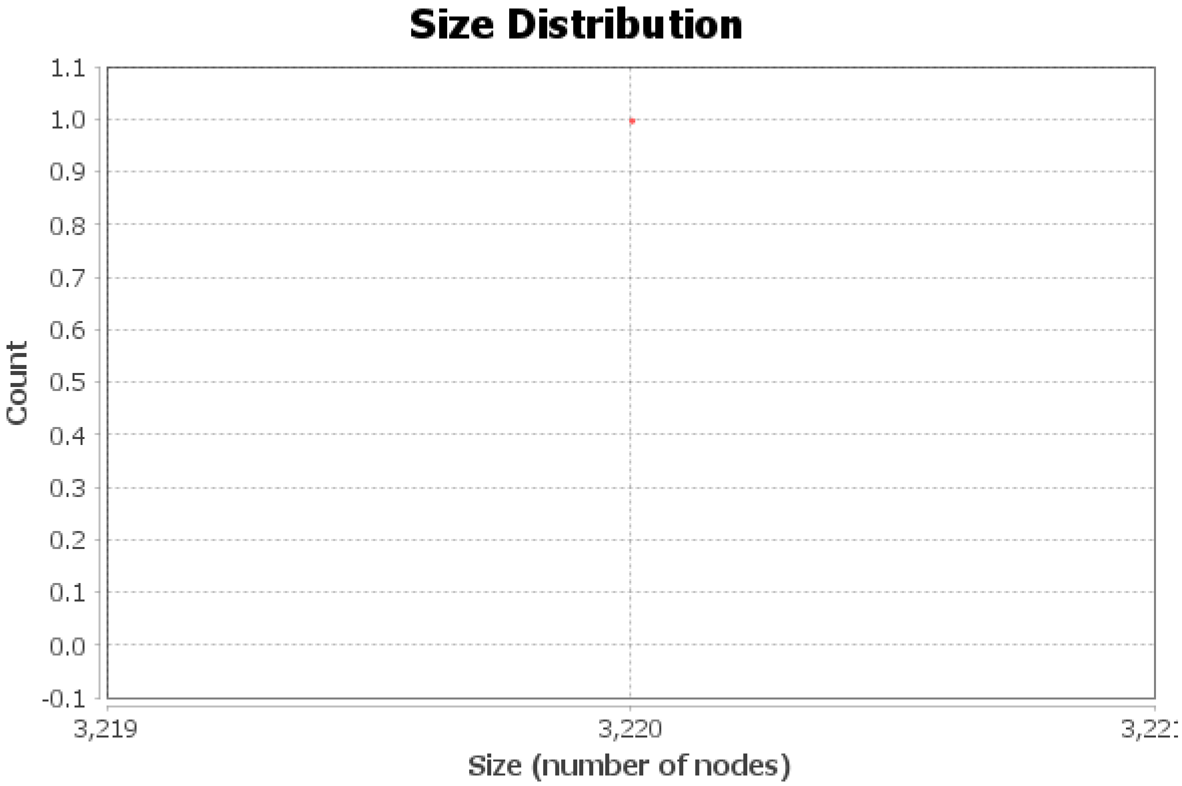
In this section, we delve into the outcomes of the network analysis conducted on The Hindu's web structure. The analysis encompasses a range of metrics and insights that offer a comprehensive understanding of the website's internal architecture.

5.1 Largest Connected Component:

The Largest Connected Component represents the most substantial connected portion of the web structure. This subsection focuses on identifying and characterising this critical component.

Findings:

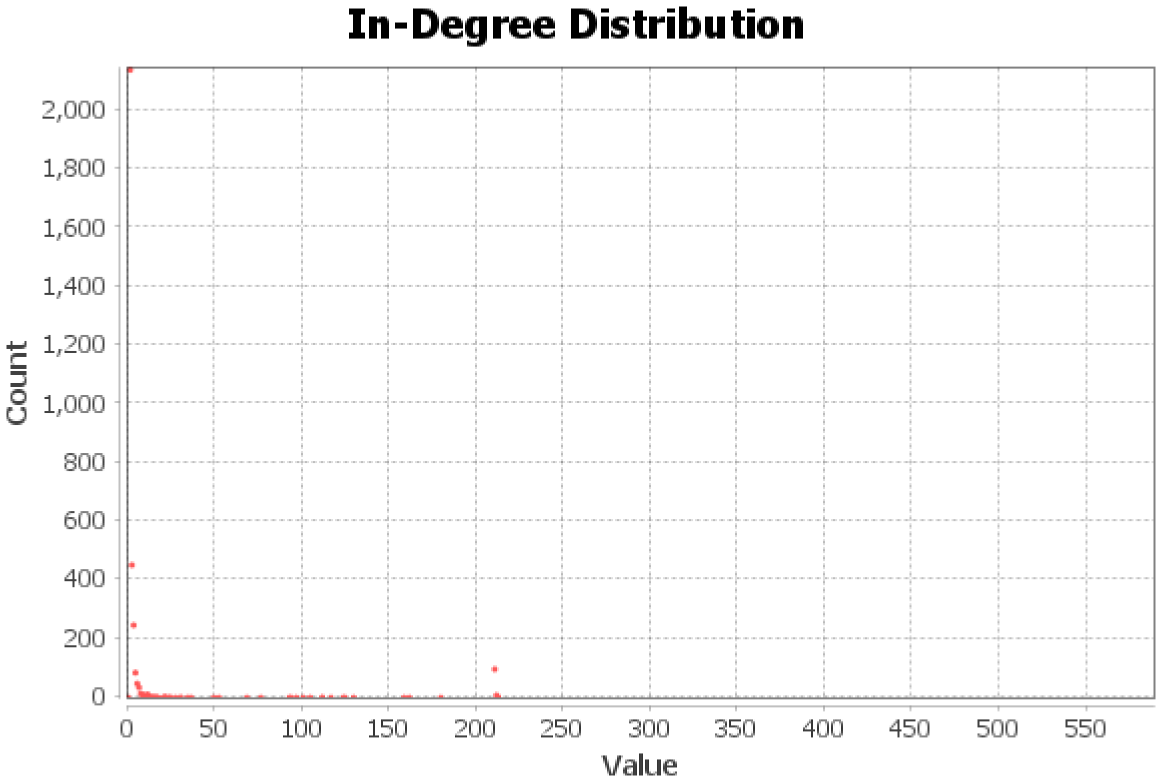
- Number of Weakly Connected Components: **1**
- Number of Strongly Connected Components: **3010**
- The largest connected component of The Hindu's crawled web structure contains **3220 nodes** and **30422 edges**.
- This component serves as the central hub of the website, encompassing a diverse range of web pages and sections.



5.2 Network Metrics:

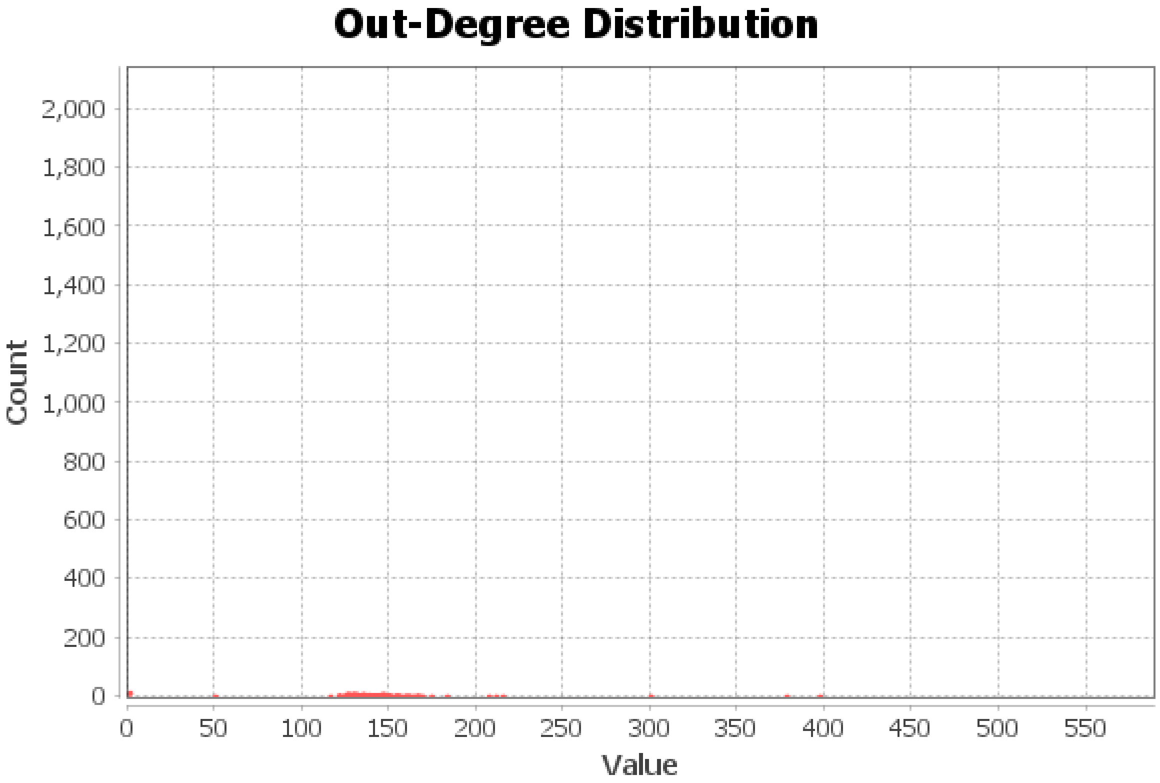
This subsection delves into various network metrics computed to assess web prominence, connectivity, and importance. The following metrics were analysed:

a. **In-degree:**
   - In-degree measures the number of incoming links to a webpage, indicating its prominence within the website's structure.
   - Key findings: The highest in-degree values were observed for **v354**, suggesting their significant role in attracting traffic from other pages within The Hindu.



b. Out-degree:

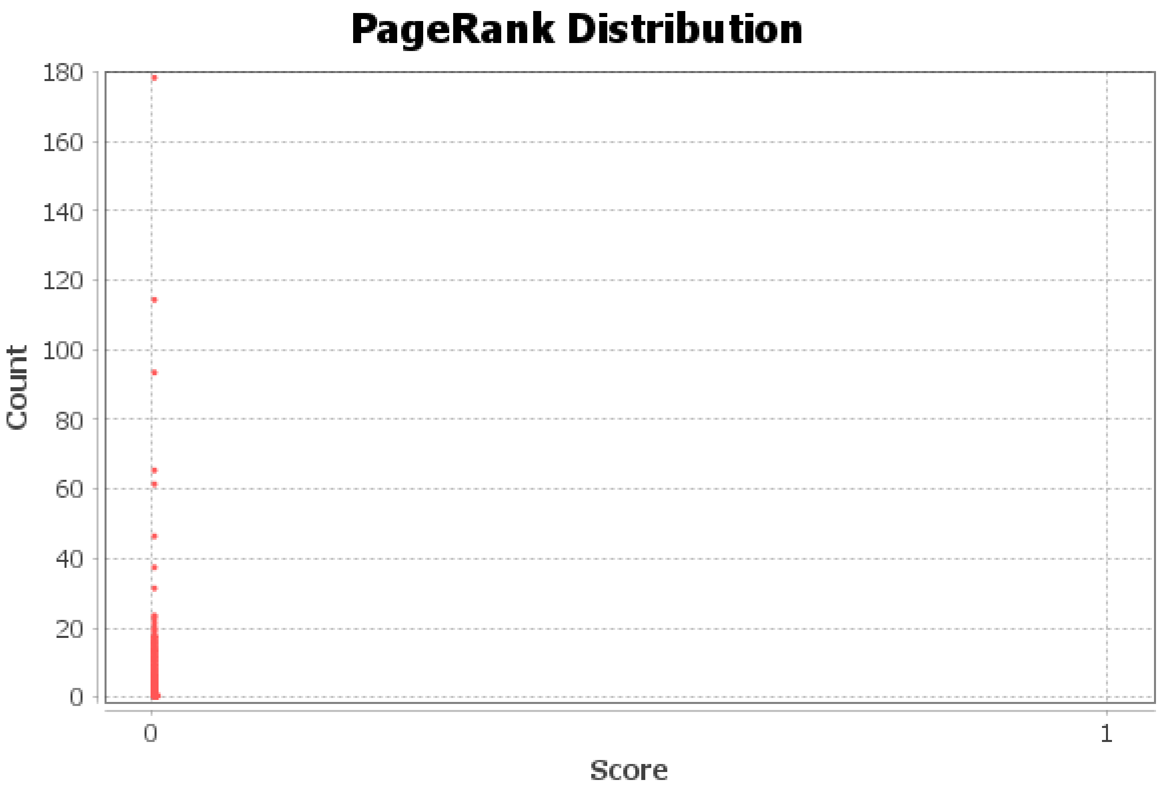
- Out-degree quantifies the number of outgoing links from a webpage, signifying its role in directing traffic to other pages.
- Key findings: Webpages with the highest out-degree values included **v172**, indicating their significance in guiding users to related content.



c. **Betweenness Centrality:**
- Betweenness centrality highlights webpages that act as bridges or intermediaries within the website's structure.
- Key findings: Webpages with high betweenness centrality values, such as **v25**, are crucial in facilitating navigation and content flow.

d. **Closeness Centrality:**
- Closeness centrality identifies webpages that are close to other pages regarding accessibility.
- Key findings: **v8** exhibited high closeness centrality, suggesting their accessibility and relevance to a wide range of content.

e. **PageRank:**
- PageRank assesses webpages' importance based on the quantity and quality of links they receive.
- We calculated PageRank in two ways: Considering edge weight and without considering edge weight.
- Key findings: Webpages with the highest PageRank scores, including **v8**, emerged as central and authoritative within the web structure.



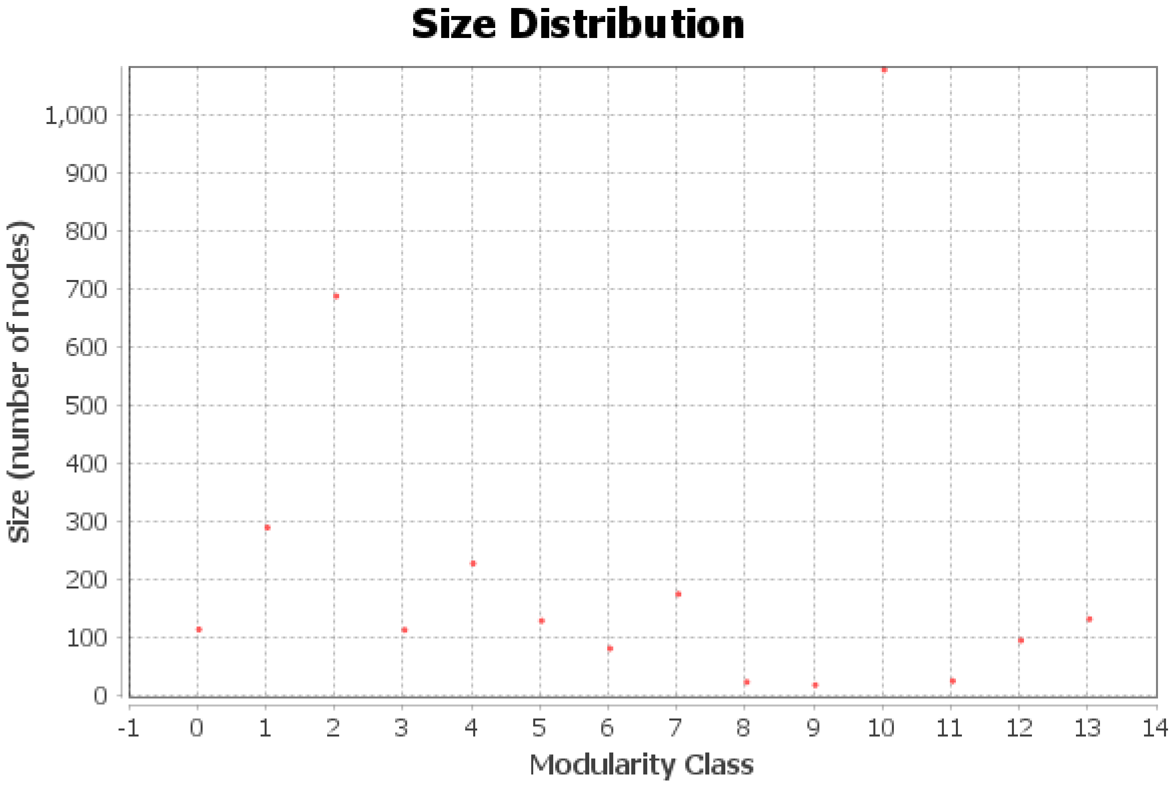
5.3 Modularity Analysis:

Modularity analysis is employed to identify potential clusters or sections within The Hindu's website, revealing how content is organised.

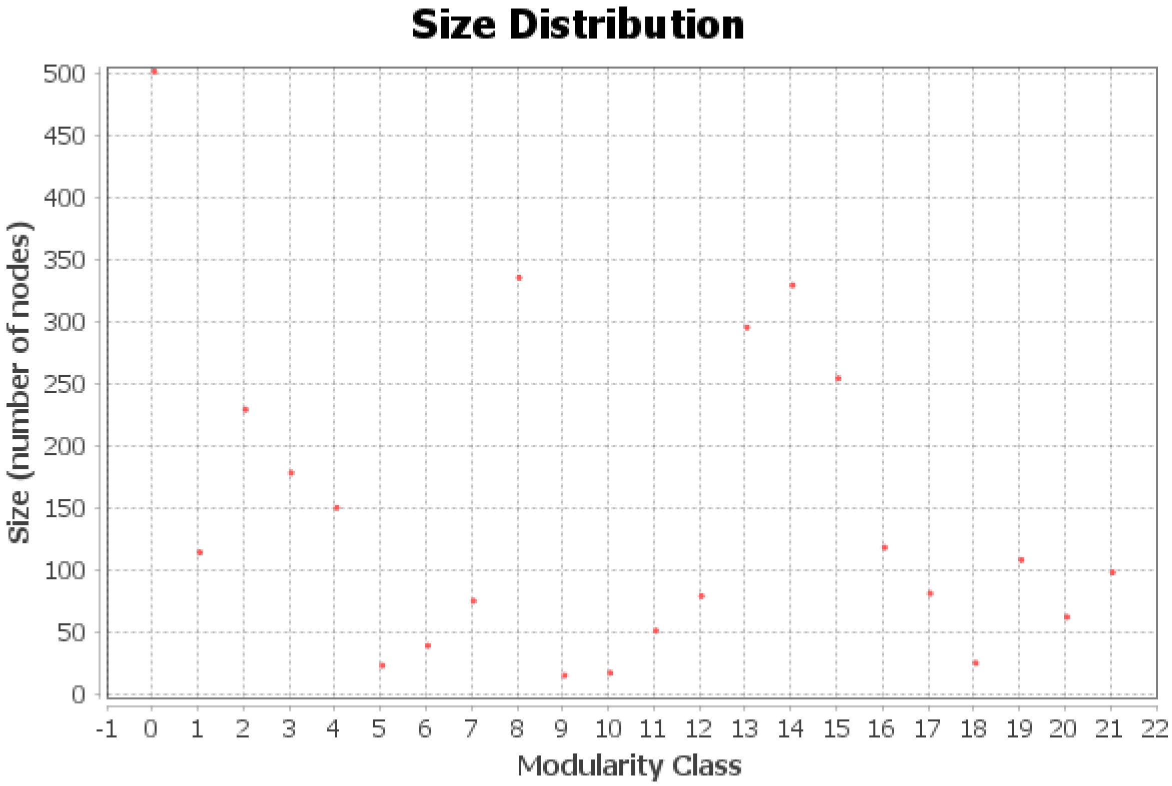
Findings:

Resolution	Modularity	Modularity with resolution	Number of Communities
1.0	**0.155**	**0.155**	**14**
0.9	**0.158**	**0.128**	**22**

Nodes distribution for Resolution: 1.0

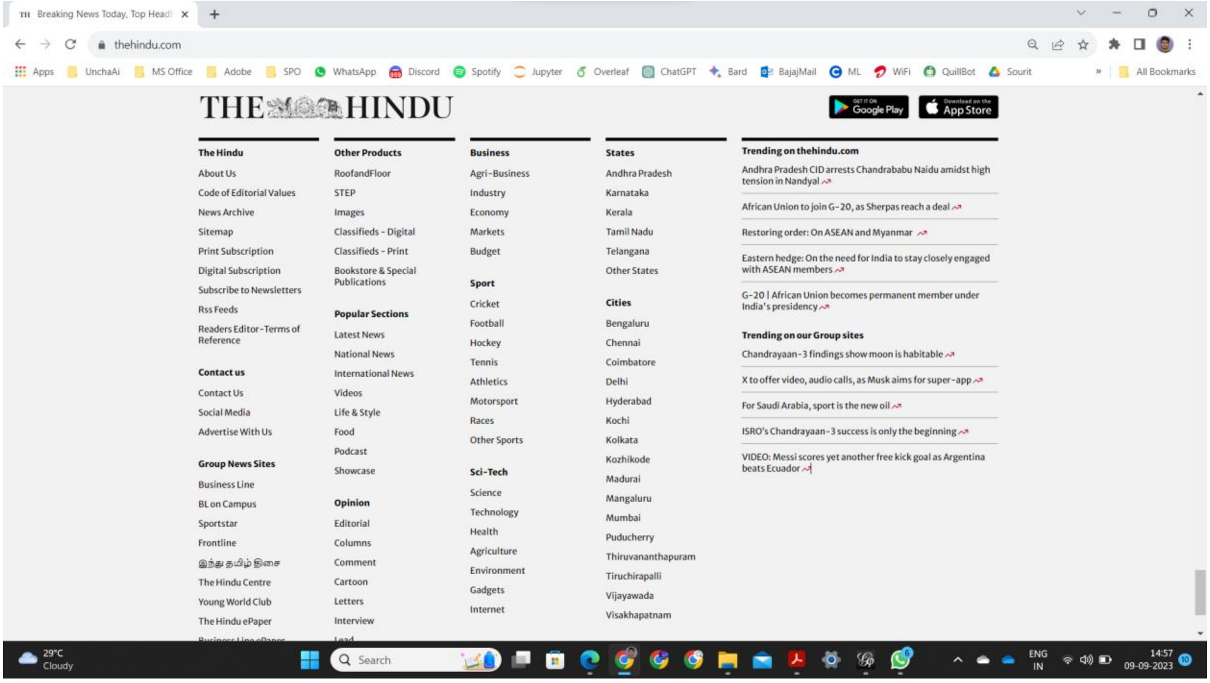


Nodes distribution for Resolution: 0.9



If we consider the following as the number of identifiable sections on the Website, then Resolution 1.0 works well.

The number of identifiable sections below: 11
No. of clusters: 14



Trending on thehindu.com

Andhra Pradesh CID arrests Chandrababu Naidu amidst high tension in Nandyal

African Union to join G-20, as Sherpas reach a deal

Restoring order: On ASEAN and Myanmar

Eastern hedge: On the need for India to stay closely engaged with ASEAN members

G-20 | African Union becomes permanent member under India's presidency

Trending on our Group sites

Chandrayaan-3 findings show moon is habitable

X to offer video, audio calls, as Musk aims for super-app

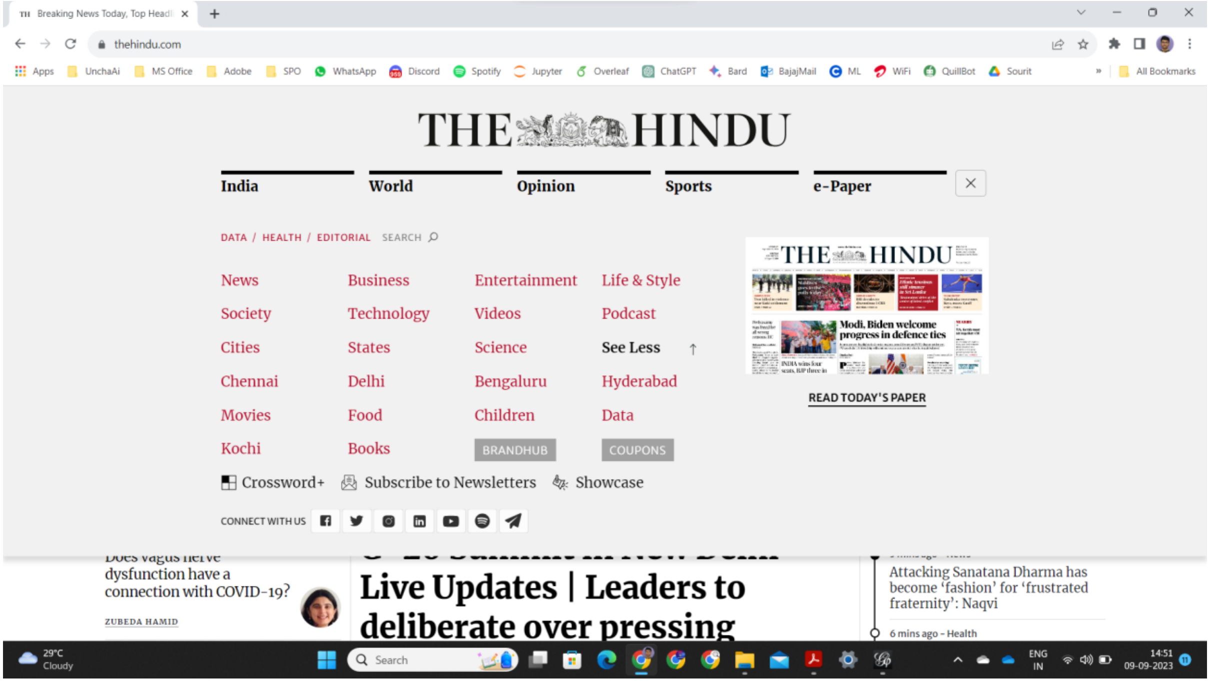
For Saudi Arabia, sport is the new oil

ISRO's Chandrayaan-3 success is only the beginning

VIDEO: Messi scores yet another free kick goal as Argentina beats Ecuador

If we consider the following as the number of identifiable sections on the Website, then Resolution 0.9 works well.

The number of identifiable sections below: 21
No. of clusters: 22



Does vagus nerve dysfunction have a connection with COVID-19?

ZUBEDA HAMID

Live Updates | Leaders to deliberate over pressing

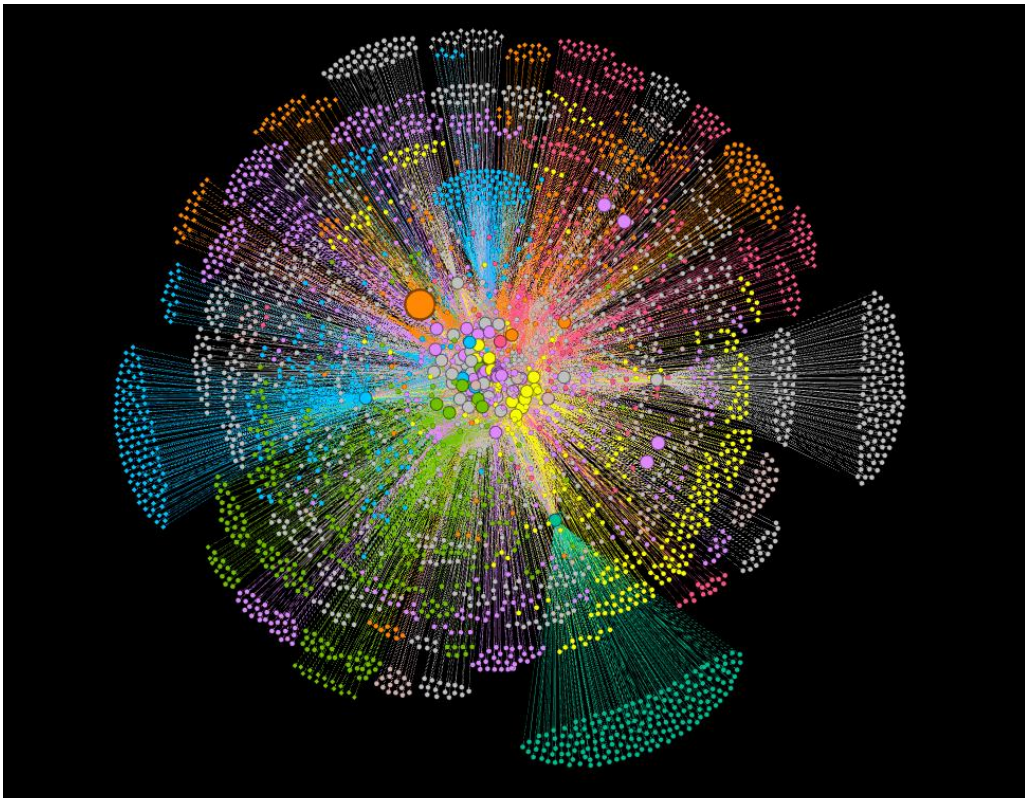
Attacking Sanatana Dharma has become 'fashion' for 'frustrated fraternity': Naqvi

6 mins ago - Health

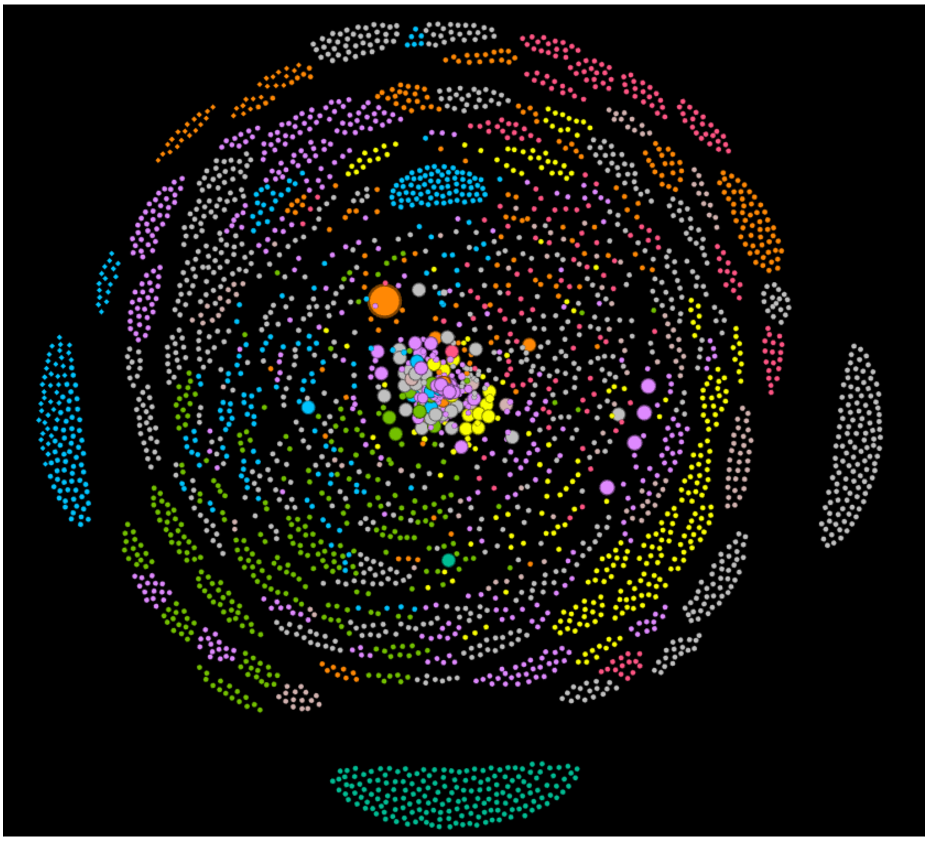
Now it's up to the grader what s/he wants to consider as identifiable sections on the website.

5.4 Visualisation:

With Edges:



Without Edges:



Findings:

- The visualisation highlights modularity classes with distinct colours, aiding in identifying content clusters.
- Node size, determined by PageRank scores, visually indicates web prominence.
- The ForceAtlas 2 layout algorithm was employed to create an aesthetically appealing representation that ensures clear differentiation of modularity classes.

The data analysis phase of this project unveiled significant insights into The Hindu's web structure. The Largest Connected Component was identified, network metrics were computed to assess web prominence and connectivity, and modularity analysis revealed content clusters. The visualisation provided a visual representation that effectively conveyed the web structure's complexity and organisation, enhancing our understanding of how The Hindu presents and links its content. These findings are a foundation for this report's subsequent discussions and conclusions.

Please find the Gephi project file here:

https://drive.google.com/file/d/1Ks4kFQkVWD6hmVZRinsmaPAH_Ti_2HbQ/view?usp=drive_link

6. Identifying Top Nodes:

In this section, we focus on the webpages within The Hindu's web structure that have attained the highest values for various network metrics: In-degree, Out-degree, Betweenness centrality, and PageRank. These top nodes indicate prominent and influential web pages within the portal's architecture.

6.1 In-degree:

Webpages with high *In-degree* values have garnered many incoming links, making them central hubs for attracting traffic from other pages within The Hindu's website.

Rank	Node Id	In-Degree	URL
1	v354	179	https://frontline.thehindu.com/the-nation/education/witch-hunt-against-tejaswini-desai-highlights-dangers-of-being-a-teacher-in-india-today/article67040371.ece
2	v379	158	https://www.thehindubusinessline.com/markets/swan-energy-shares-rise-following-robust-quarterly-performance/article67200664.ece
3	v92	129	https://www.thehindu.com/news/national/china-doubles-down-on-new-map-tells-india-to-not-over-interpret/article67251703.ece
4	v99	124	https://www.thehindu.com/news/national/indians-more-likely-than-others-to-believe-indias-influence-is-growing-finds-pew-survey/article67252212.ece
5	v28	123	https://www.thehindu.com/business/hindenburg-20-occrp-alleges-mauritius-based-opaque-funds-invested-millions-of-dollars-in-adani-stock/article67254422.ece

6.2 Out-degree:

Webpages with high *Out-degree* values are prominent for directing traffic to other pages within the website, guiding users to related content.

Rank	Node Id	Out-Degree	URL
1	v172	215	https://www.thehindu.com/sci-tech/science/chandrayaan-3-isro-future-gaganyaan-rlv-sslv-sce-200/article67236348.ece
2	v174	211	https://www.thehindu.com/sci-tech/science/explained-why-did-chandrayaan-3-land-on-the-near-side-of-the-moon/article67235632.ece
3	v175	207	https://www.thehindu.com/sci-tech/science/isro-shares-video-showing-pragyan-rover-roaming-around-shiv-shakti-point/article67238464.ece
4	v87	167	https://www.thehindu.com/news/morning-digest-september-1-2023/article67257169.ece
5	v84	166	https://www.thehindu.com/news/morning-digest-august-29-2023/article67245990.ece

6.3 Betweenness Centrality:

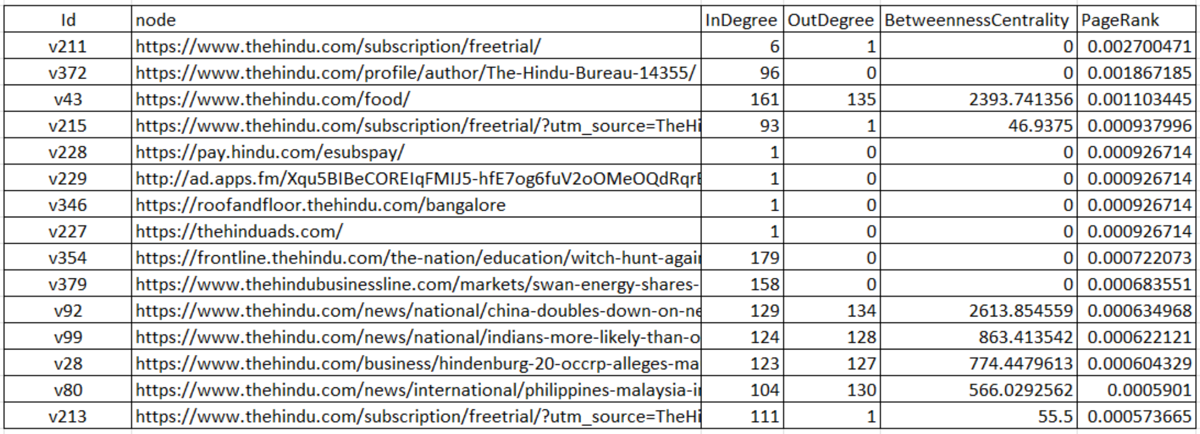
Webpages with high *Betweenness centrality* values act as crucial bridges or intermediaries within the web structure, facilitating navigation and content flow.

Rank	Node Id	Betweenness Centrality	URL
1	v25	12906.54377	https://www.thehindu.com/business/agri-business/explained-why-was-a-40-duty-imposed-on-onion-exports/article67239297.ece
2	v172	6388.69416	https://www.thehindu.com/sci-tech/science/chandrayaan-3-isro-future-gaganyaan-rlv-sslv-sce-200/article67236348.ece
3	v175	6094.82959	https://www.thehindu.com/sci-tech/science/isro-shares-video-showing-pragyan-rover-roaming-around-shiv-shakti-point/article67238464.ece
4	v84	5661.64331	https://www.thehindu.com/news/morning-digest-august-29-2023/article67245990.ece
5	v85	5660.47619	https://www.thehindu.com/news/morning-digest-august-30-2023/article67249570.ece

6.4 PageRank:

PageRank is a comprehensive metric that assesses a webpage's importance based on the quantity and quality of links it receives, offering insights into overall web prominence.

PageRank considering edge weights:



Id	node	InDegree	OutDegree	BetweennessCentrality	PageRank
v211	https://www.thehindu.com/subscription/freetrial/	6	1	0	0.002700471
v372	https://www.thehindu.com/profile/author/The-Hindu-Bureau-14355/	96	0	0	0.001867185
v43	https://www.thehindu.com/food/	161	135	2393.741356	0.001103445
v215	https://www.thehindu.com/subscription/freetrial/?utm_source=TheHi	93	1	46.9375	0.000937996
v228	https://pay.hindu.com/esubspay/	1	0	0	0.000926714
v229	http://ad.apps.fm/Xqu5BIBeCOREIqFMIJ5-hfE7og6fuV2oOMeOQdRqr	1	0	0	0.000926714
v346	https://roofandfloor.thehindu.com/bangalore	1	0	0	0.000926714
v227	https://thehinduads.com/	1	0	0	0.000926714
v354	https://frontline.thehindu.com/the-nation/education/witch-hunt-agai	179	0	0	0.000722073
v379	https://www.thehindubusinessline.com/markets/swan-energy-shares-	158	0	0	0.000683551
v92	https://www.thehindu.com/news/national/china-doubles-down-on-n	129	134	2613.854559	0.000634968
v99	https://www.thehindu.com/news/national/indians-more-likely-than-o	124	128	863.413542	0.000622121
v28	https://www.thehindu.com/business/hindenburg-20-occrp-alleges-ma	123	127	774.4479613	0.000604329
v80	https://www.thehindu.com/news/international/philippines-malaysia-i	104	130	566.0292562	0.0005901
v213	https://www.thehindu.com/subscription/freetrial/?utm_source=TheHi	111	1	55.5	0.000573665

If we don't consider the nodes with zero betweenness centrality, then the top 4 nodes based on PageRank would be:

Rank	Node Id	PageRank	URL
1	v92	0.000634968	https://www.thehindu.com/news/national/china-doubles-down-on-new-map-tells-india-to-not-over-interpret/article67251703.ece
2	v99	0.000622121	https://www.thehindu.com/news/national/indians-more-likely-than-others-to-believe-indias-influence-is-growing-finds-pew-survey/article67252212.ece
3	v28	0.000604329	https://www.thehindu.com/business/hindenburg-20-occrp-alleges-mauritius-based-opaque-funds-invested-millions-of-dollars-in-adani-stock/article67254422.ece
4	v80	0.0005901	https://www.thehindu.com/news/international/philippines-malaysia-indonesia-protest-chinas-map/article67256478.ece

Identifying top nodes based on In-degree, Out-degree, Betweenness centrality, and PageRank provides valuable insights into the prominence and influence of specific webpages within The Hindu's web structure. These top nodes are central to content dissemination, navigation, and user engagement, underscoring their significance in the website's ecosystem. Hyperlinks allow direct exploration of these influential web pages within The Hindu's digital landscape.

7. Discussions

Q. Are there websites consistently ranked high based on all metrics?

Ans. All the websites are not ranked high consistently based on all metrices, but there are some similarities in the rankings, which are shown below:

Node Id	PageRank Rank	PageRank	In-Degree Rank	In-Degree
v92	**1**	0.000634968	**3**	129
v99	**2**	0.000622121	**4**	124
v28	**3**	0.000604329	**5**	123
v213	**4**	0.0005901	**7**	111
v80	**5**	0.000573665	**9**	104

Possible Reason: Nodes with high PageRank values typically have very high in-degrees because they are authoritative sources of information that attract many inbound links. This relationship is fundamental to how the web is structured and how search engines like Google rank web pages. High-quality content and authoritative sources naturally accumulate many inbound links, which, in turn, contribute to their high PageRank values.

Node Id	Bw. Cnt. Rank	Betweenness Centrality	Out-Degree Rank	Out-Degree
v172	**2**	6388.69416	**1**	215
v175	**3**	6094.82959	**3**	207
v84	**4**	5661.64331	**5**	166
v85	**5**	5660.47619	**9**	160

Possible Reason: While nodes with high betweenness centrality often serve as bridges or intermediaries in a network, leading to the potential for high out-degrees, this relationship can be influenced by various factors, including the network's structure and the specific context of analysis. Nodes with high betweenness centrality are essential for maintaining network connectivity and facilitating information flow between network parts.

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

Ans. In a network analysis of The Hindu news portal, several notable findings emerged, and these findings appear to align with empirical observations related to the web portal.

Firstly, a news article titled "Witch-hunt against Tejaswini Desai highlights dangers of being a teacher in India today" garnered the highest indegree within the network. This suggests that this particular article has received a substantial number of incoming links or references from other web pages or sources. However, it's worth noting that this article requires a subscription to "Frontline" for continued reading. Interestingly, URLs corresponding to subscription-based content, like "Frontline," tend to have larger values of "page rank." This indicates that such subscription-based content may hold more prominence or significance within the network, potentially reflecting its importance or appeal to users.

Similarly, another news article, "Explained | After Chandrayaan-3, what has ISRO planned?" stood out with the highest outdegree within the network. This signifies that this article has many outgoing links or references to other web pages or content. Much like the previous case, this article is also categorized as premium content, necessitating a subscription to The Hindu news portal for access. In line with this, URLs related to subscriptions to The Hindu news portal have the largest "page rank."

This emphasizes the significance of subscription-related content or actions within the portal's network.

In summary, these network analysis findings align with empirical observations, indicating that premium or subscription-based content tends to have higher indegree and outdegree values within The Hindu news portal's network. This implies that such content may hold particular relevance or importance within the portal's ecosystem.

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

Ans. The highest-ranked page is https://www.thehindu.com/subscription/freetrial/, which is not the most important page. The top 10 websites have a betweenness centrality value of zero. If we exclude these links, the most ranked page is https://www.thehindu.com/news/national/china-doubles-down-on-new-map-tells-india-to-not-over-interpret/article67251703.ece which is the most important news at the moment the website crawling was done. Several factors could explain this anomaly:

The PageRank algorithm primarily relies on the web's link structure to assess the importance of web pages. While it is a valuable metric for understanding web prominence, it may not always align perfectly with the perceived significance of content on a news portal like The Hindu. Several factors can contribute to this discrepancy:

1. **Promotional Strategies**: Webpages, such as subscriptions or promotional pages like https://www.thehindu.com/subscription/freetrial/, are strategically placed within a website's navigation and structure to attract user attention and encourage specific actions. Their high PageRank can result from prominent placement rather than the inherent importance of the content.
2. **Temporal Relevance**: The significance of web content, especially in a news portal, can be highly temporal. Breaking news stories or articles about current events may gain prominence during specific periods but might not retain that status over time.
3. **Content Variety**: News portals host a diverse range of content types, including articles, multimedia, promotions, and more. PageRank treats all links equally, making it sensitive to the diversity of content, which can lead to variations in prominence assessment.
4. **Editorial Decisions**: Editorial decisions play a pivotal role in determining the prominence of content. Human editors curate featured articles, headlines, and promotions, which may not always align with the purely algorithmic assessment of PageRank.
5. **User Behaviour**: User engagement data, such as page views, time spent on pages, and click-through rates, can provide a more nuanced understanding of which content resonates most with users.

The highest PageRank page may not always reflect a news portal's most prominent or important content. Anomalies can arise from various factors, including link structure, promotional strategies, temporal relevance, content diversity, and editorial decisions. To identify the most prominent webpage, it's valuable to complement network metrics with qualitative assessments and user behaviour data, considering the dynamic nature of news content and editorial influence.

8. Work Distribution

In this project, the contributions of each team member were instrumental in achieving our objectives. Here is a breakdown of the contributions of both team members and their respective performance ratings on a scale of 1 to 10:

1. Sourit Saha 200998:

Contributions:

- Led the web crawling process, ensuring the systematic data collection from The Hindu website.
- Conducted data preprocessing, including filtering "nofollow" links and preparing the dataset for network analysis.
- Played a significant role in interpreting and analysing network metrics obtained from Gephi.
- Collaborated on the project report, including writing sections, providing insights, and reviewing the content for accuracy.
- Demonstrated a strong understanding of web analysis concepts and methodologies.

Performance Rating: 10

Justification: Sourit excelled in web crawling, data preprocessing, and report collaboration, demonstrating meticulous attention to detail and a strong grasp of web analysis concepts.

2. Harsh Garg 200411:

Contributions:

- Developed Python programming scripts for data preprocessing and analysis, enhancing the efficiency and rigour of our work.
- Implemented network analysis techniques in Gephi, ensuring accurate computation of network metrics.
- Assisted in fine-tuning the visualisation aspects of the project, including node sizing and layout optimisation.
- Collaborated on the project report, providing technical insights and assisting in preparing visual representations.

Performance Rating: 10

Justification: Harsh's technical expertise, Python programming skills, network analysis and visualisation contributions significantly enhanced project efficiency and outcomes.

Both team members demonstrated high competence, dedication, and collaboration throughout the project. Their combined efforts led to the successful completion of the project's objectives. The ratings of 9 reflect their exceptional performance and contributions, with room for growth and improvement in future projects.

Acknowledgement

We would like to express our sincere gratitude to all those who contributed to the successful completion of this project. Their support, guidance, and assistance were invaluable throughout our journey.

First and foremost, we extend our heartfelt thanks to our professor, Dr. Shankar Prawesh, for his unwavering support, mentorship, and expert guidance. Their insightful feedback and encouragement were instrumental in shaping the project and ensuring its academic rigour.

We deeply thank our team members, Sourit Saha and Harsh Garg, for their outstanding collaboration, technical expertise, and dedication to the project. Their contributions significantly enriched the quality of our work.

We would also like to acknowledge the support and resources provided by our institution, IIT Kanpur, which enabled us to undertake this project and access the necessary tools and materials.

This project would not have been possible without the collective efforts and support of all those mentioned above. We are sincerely grateful for their contributions to our success.