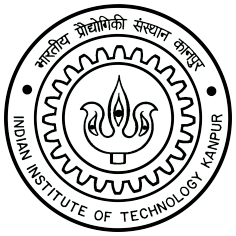

Link Collection and Graph Analysis using Screaming Frog and Gephi

Amipriya Anand (220122) Harsh Nirmal (220431)

MBA749M - Social Media Analytics

Instructor: *Prof.* Shankar Prawesh

September 7, 2025

Abstract

This report presents the process of hyperlink collection and network analysis using Screaming Frog SEO and Gephi. Starting with the IIT Kanpur Wikipedia page as the seed, we systematically crawled and curated web addresses to construct a directed graph. The network was analyzed using key graph algorithms such as PageRank, in-degree/out-degree distribution, and modularity clustering. Results are visualized and discussed.

1 Introduction

The objective of this project is to create a **large-scale web graph** and analyze its structure. The workflow involves:

1. Crawling hyperlinks using **Screaming Frog SEO**.
2. Cleaning and preparing the data.
3. Importing the graph into **Gephi** for visualization and analysis.

2 Methodology

2.1 Data Collection

We used Screaming Frog SEO with the IIT Kanpur Wikipedia page as the seed URL. The crawler was run iteratively to collect links and construct successive seed lists.

2.2 Data Cleaning and Preparation

From the raw extracted data, we retained only the `From` and `To` columns. Duplicates, protocols (`https://`, `http://`, `www.`), and excessively long (size > 100) URLs were removed. The final dataset was saved in CSV format.

2.3 Graph Import in Gephi

The curated dataset was imported into Gephi. The Yifan Hu layout was applied for visualization.

3 Results and Analysis

3.1 Graph Visualization

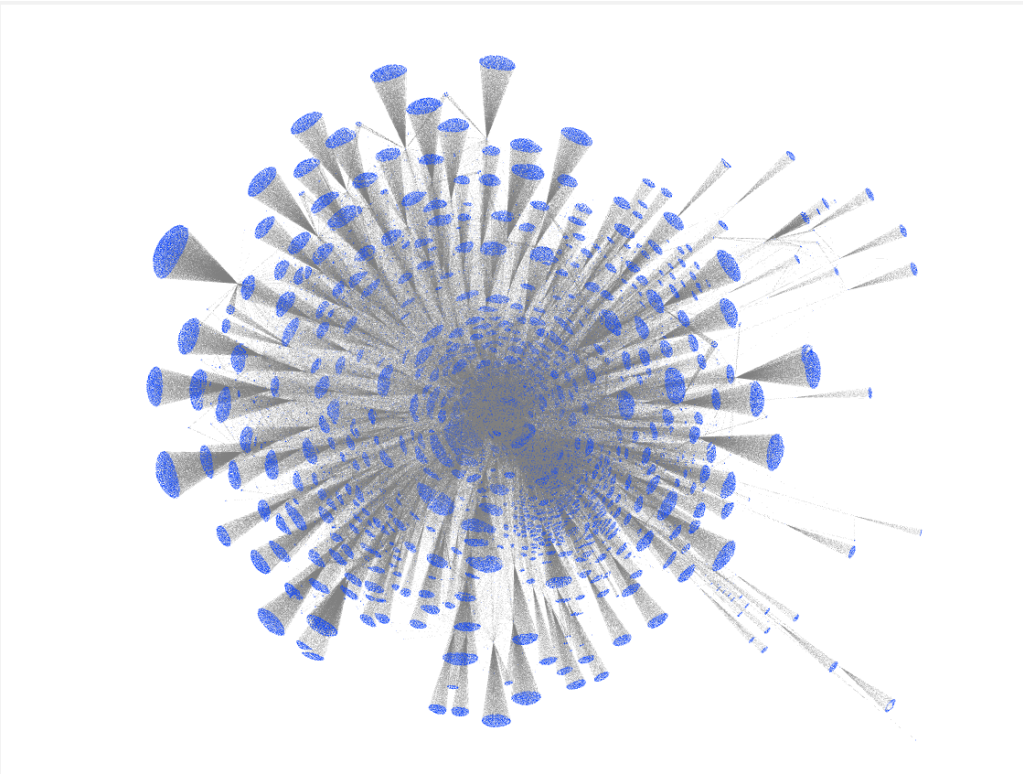

Figure 1: Web graph visualization using Yifan Hu layout (Nodes in Blue and Edges in Gray).

3.2 Number of Nodes and Edges

The imported graph contained:

- **Nodes**: 154381
- **Edges**: 302964

3.3 Top-10 Nodes by Degree

Below are the top 10 nodes reported as per In-Degree & Out-Degree:

Nodes	In-Degree
developer.wikimedia.org/	570
mediawiki.org/	570
foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Universal$_{C}ode_{o}f_{C}onduct$	569
stats.wikimedia.org/	569
foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Cookie$_{s}tatement$	569
wikimedia.org/	569
foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Privacy$_{p}olicy$	565
foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Terms$_{o}f_{U}se$	556
wikimediafoundation.org/	548
en.wikipedia.org/wiki/Main$_{P}age$	547

Table 1: Top-10 nodes ranked by In-Degree

Nodes	Out-Degree
en.wikipedia.org/wiki/New$_{Y}ork_{C}ity$	3903
en.wikipedia.org/wiki/University$_{o}f_{M}ichigan$	2770
en.wikipedia.org/wiki/India	2744
en.wikipedia.org/wiki/University$_{o}f_{J}ohannesburg$	2502
en.wikipedia.org/wiki/MIT	2222
en.wikipedia.org/wiki/University$_{o}f_{t}he_{W}itwatersrand$	2212
en.wikipedia.org/wiki/Wikipedia:Community$_{p}ortal$	2196
en.wikipedia.org/wiki/Sanskrit	2134
en.wikipedia.org/wiki/Diwali	2095

Table 2: Top-10 nodes ranked by Out-Degree

3.4 PageRank Results

Below are the top 10 nodes based on Page Rank Metric:

Nodes	PageRanks
mediawiki.org/wiki/MediaWiki	0.000025
stats.wikimedia.org/assets-v2/main.css	0.000024
developer.wikimedia.org/	0.000021
wikimedia.org/	0.000021
mediawiki.org/	0.000021
stats.wikimedia.org/	0.000021
foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Universal$_{C}ode_{o}f_{C}onduct$	0.000021
foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Cookie$_{s}tatement$	0.000021
foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Privacy$_{p}olicy$	0.000021
foundation.wikimedia.org/wiki/Special:MyLanguage/Policy:Terms$_{o}f_{U}se$	0.000021

Table 3: Top-10 nodes ranked by Out-Degree

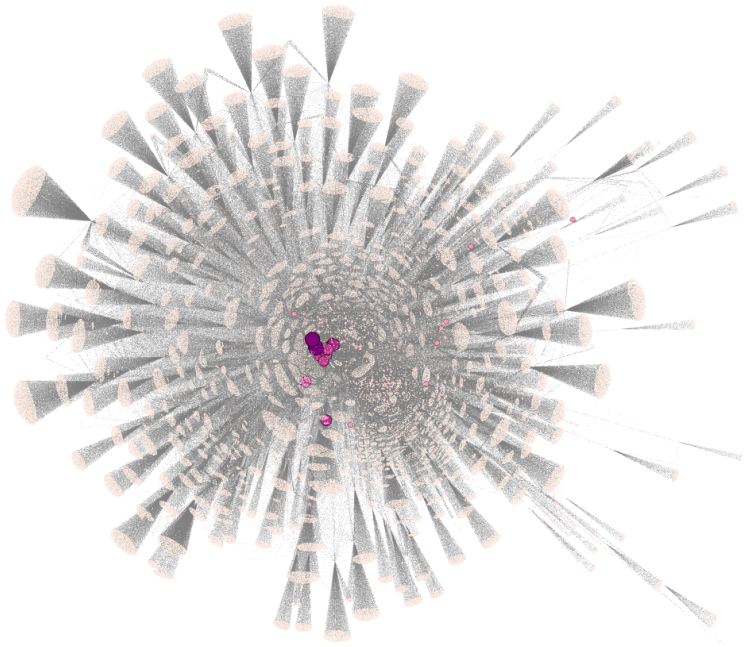

Figure 2: Graph visualization with node sizes and colour scaled by PageRank (Higher the PageRank, Darker the Pink shade).

3.5 Modularity and Clustering

The modularity algorithm was applied to detect communities. Nodes were colored by modularity class. There were **total 41 classes** detected by this algorithm.

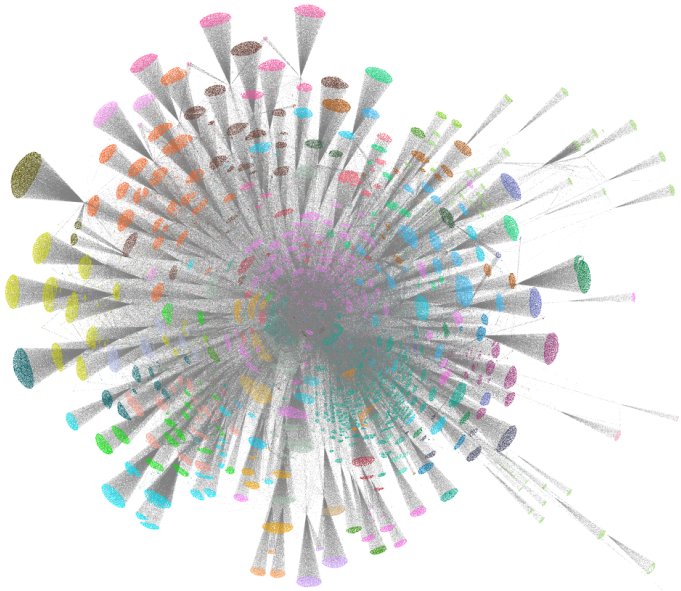

Figure 3: Clusters detected by modularity algorithm.

3.6 Largest Cluster Analysis

The largest modularity-based cluster (**modularity class = 31**) was filtered and exported.

- **Nodes**: 14626
- **Edges**: 53356

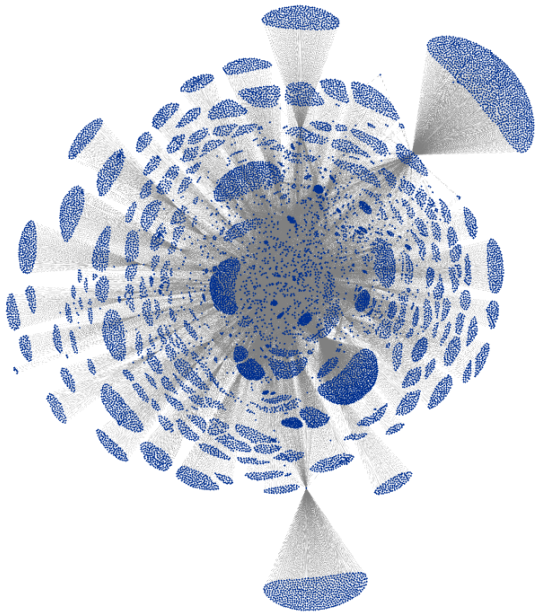

Figure 4: Visualization of the largest cluster with node labels

Interpretation: The nodes in this cluster are grouped together due to their high link density and shared topical relevance. Specifically:

- They likely represent web pages related to IIT Kanpur's academic departments, research centers, or administrative units.
- The hyperlink structure — menus, footers, and internal references—creates strong interlinking among these pages.
- The modularity algorithm detects this structural cohesion, grouping them into a distinct community.
- Due to their dense internal connectivity, the modularity algorithm classifies these pages within the same community, effectively recognizing that these nodes form a closely interrelated subnetwork.

4 Discussion

The analysis reveals structural properties of the web graph, highlighting important hubs, authorities, and communities. The largest cluster indicates highly interrelated nodes, likely belonging to closely linked domains or subtopics of IIT Kanpur.

5 Conclusion

We successfully constructed a hyperlink graph and analyzed it using centrality measures and community detection in Gephi. This approach can be extended to other domains to understand link structures and influence.

Team Contributions

- **Amipriya Anand**: Data collection and preprocessing.
- **Harsh Nirmal**: Graph import and visualization.
- **Both**: Analysis and report writing.