Analyzing Web Graph Dataset using Ranking, Clustering and Centrality Metrics

Nabila Shahnaz Khan
University of Central Florida

Abstract

Network analysis is very useful in deep understanding of the structure and relationship among the entities of a connected network. Using different network analysis algorithms and centrality measures, important characteristic features of a network can be derived. In this paper, a portion of the well-known dataset of Stanford Web graph has been analyzed using different network centrality measures and Web page ranking algorithms to identify more significant web pages in the web graph. Later, network clustering was performed to partition the web graph into more densely connected clusters. According to the analysis results, it has been seen that some web pages can be identified as more important based on different ranking scores and centrality metrics. Also, the clustering result shows that the network isn't a uniformly dense network, rather it shows community structure property like most other real life networks.

Index Terms

Network Analysis, Network Centrality, Page Rank Algorithm, Weighted Page Rank Algorithm, HITS, MCL Clustering

I. Introduction

Network can be defined as a collection of objects connected to each other based on certain similarity of features. Network analysis is considered to be a set of integrated techniques used to analyze the network for identifying the type and reasoning of relations formed among the objects in a network. In order to understand a network structure better and discover the inter-relations between objects, the importance of network analysis is tremendous. Almost all the existing networks can be represented using Graph data structure where the objects are represented as nodes and the relation among them is shown using edges. Web graph is a kind of graph based network where the web pages are represented as nodes and the hyperlinks are represented using edges. The goal for analyzing Web graph is to study the link patterns emerging between documents and web sites on the World Wide Web [1]. Web graph analysis is of immense importance for getting the most relevant and valuable information and knowledge from the huge pool of data contained in the internet. For studying Web graph networks, different types of analysis techniques and algorithms are available till date. Some highly used techniques are network centrality analysis, network clustering etc. Apart from that, different algorithms such as PageRank (PR), Weighted PageRank (WPR), Hyperlink-Induced Topic Search (HITS), Spamming Resistant Expertise Analysis and Ranking (SPEAR) etc. have been developed for ranking the web pages based on their importance and connectivity [2].

In this work, in order to learn and understand web graph network analysis better, a segment of the well-known Stanford Web Graph Dataset has been analyzed using techniques like Network centrality, Network clustering and algorithms such as PageRank, Weighted PageRank, HITS. Here, Section II presents the literature review of web graph analysis while Section III gives an idea about the overall goal of this work. Section IV introduces the dataset used for this project. Methodology and analysis of the result have been elaborately discussed in Section V. Finally, Section VI concludes the paper.

II. Background

From the mathematical and data structural point of view, different real life problems can be represented as graph based network model such as transportation networks, social networks, biological networks, utility networks and thus can be solved using graph analysis techniques and algorithms. As a result, network analysis has become a very integral part of data analysis and mining. The World Wide Web is a great source of data and information. According to a study taken place in the early 2000s, it contained over a billion web pages with near about seven billion hyperlinks [3]. By now, this number has increased enormously beyond all expectations. According to statistical research by WorldWideWebSize, as of 2020, the web contains over 6 billion indexed web pages [4]. Different Web graph based research works have been done from time-to-time to improve the quality of web search and information retrieval [5, 6].

An important aspect of Web Graph analysis is identifying the most important nodes which uses techniques like network centrality. The concept of network centrality was first introduced in the field of Social Network Analysis (SAN) [7]. Another important analysis criteria is figuring out most related and linked web pages which uses different network clustering approaches. The clustering approach uses different algorithms such as Highly Connected Subgraph (HCS), Restricted neighborhood search clustering (RNSC), Molecular Complex Detection (MCODE), Markov Cluster Algorithm (MCL) etc. One more important criteria regularly used by web search engines is ranking the webpages based on their importance and popularity. For this, different type of web page ranking algorithms have been introduced. Among these, PageRank is one of the most well-established algorithms and it is still being used by the most popular search engine Google [8]. Weighted PageRank is an extension of the PageRank algorithm, which takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages [9]. Another very popular algorithm is Hyperlink-Induced Topic Search (HITS) [10] which rank web-pages based on Hub and Authority scores. In this work, I have tried to use all these above mentioned approaches to analyze a fraction of the Stanford Web graph dataset and later compared those results to see their ranking similarities and variance.

III. Problem Statement

Web graph analysis plays an important role in making web data mining and valuable information gathering easier and faster. It is also used to improve the quality of web page ranking and search engines. In this work I have evaluated a fraction of the well-known Stanford web graph dataset in order to retrive significant analytical information. The analysis was done using the following approaches: (1) Analyzing Web graph characteristics using Degree Distribution and Degree of Separation. (2) Identifying the most important web pages in the network based on the concept of network centrality. (3) Figuring out the ranking of the web pages based on some state-of-the-art algorithms such as PageRank, Weighted PageRank and HITS. (4) Finding the highly related and connected structures within the web graph using Markov Cluster Algorithm.

IV. Dataset

A small portion of the famous Stanford Web graph dataset "web-Stanford" collected by SNAP [11] has been used in this work. This dataset has been collected in the year of 2002. It consists of a total of 281,903 nodes and 2,312,497 directed edges. Here, the nodes represent web pages from Stanford University and the edges represent the hyperlinks between the pages. Analyzing a network this big requires huge amount of computational time. So, instead of analyzing the whole network, here I have worked on a small portion of the network containing 5,094 nodes and 11,996 directed edges.

V. Methodology and Result Analysis

A. Analyzing Characteristics

1) Degree Distribution: For any degree k, the number of nodes having degree k is known as the frequency f(k) of degree k. The degree distribution P(k) of a network is defined to be the fraction of nodes in the network with degree k. For directed graph, the degree frequency and degree distribution of inwards and outwards edges are considered separately. Here, Figure 1 shows the indegree and outdegree frequency, f(k) for the given input graph while Figure 2 shows the the indegree and outdegree distribution probability, p(k). From the Figure 1, it can be seen that the maximum indegree and outdegree are 231 and 211, respectively, which is pretty high. However, the minimum degree is 0 for both indegree and outdegree with a very high probability distribution. The reason is, as a fraction of the graph has been considered for the analysis, there are lots of nodes in the input dataset whose edges exists in the original dataset but haven't been included in the smaller input portion. Other than that, it can be seen that degree range 1-10 has comparatively high distribution for both inward and outward edges. In spite of providing a very small portion of the actual dataset, the degree frequency and distribution seems high. So, it can be assumed that the original large network dataset is very dense. Also, like most of the real life networks, the degree distribution here also seems to follow the power law distribution. Here, most of the nodes have a degree ranging from 1-5 (0 excluded) and very few nodes have high degree.

2) Degree of Separation: Degree of separation of any two nodes represents the distance between those two nodes. This distance is basically the number of edges required to traverse from one node to another. This term is more commonly used in Social Network Analysis (SAN). In 1929, Hungarian writer Frigyes Karinthy introduced the concept of "Six degrees of separation". This basically establised the idea that each person is at most six steps away from being connected to any other person in a social network. This concept later have been proved by different experiments and became very popular [12, 13]. Degree of separation of a graph basically gives an idea about how closely connected are the nodes in that graph. The different degree of separation analysis values for the given input graph has been shown in the Table I. Here, the maximum degree of separation is 13 which means that the node pairs with max distance has 13 edges between them. This is also known as the diameter of the network. The minimum distance is 1 which is obvious as there will be at least two nodes in a graph which are connected to each other though single edge. The median, mode and average degree of separation for the input graph is 5,4 and 5 respectively. Which means, most of the nodes in the graph has a distance of 4-5 edges between them.

B. Identifying Important Web Pages Using Network Centrality

Network centrality tries to measure which nodes are most central. In other words, which nodes are most important based on their positions. The definition of importance of nodes may vary depending on the types of networks. In case of web graphs, mostly the centrality of web page points out the web pages which are closely connected to other web pages and contain more information. Centrality check is important for web graphs cause it measures the trustworthiness of web pages, hence decreasing the risk of being spammed. The basic network centrality metrics used in this project have been described below:

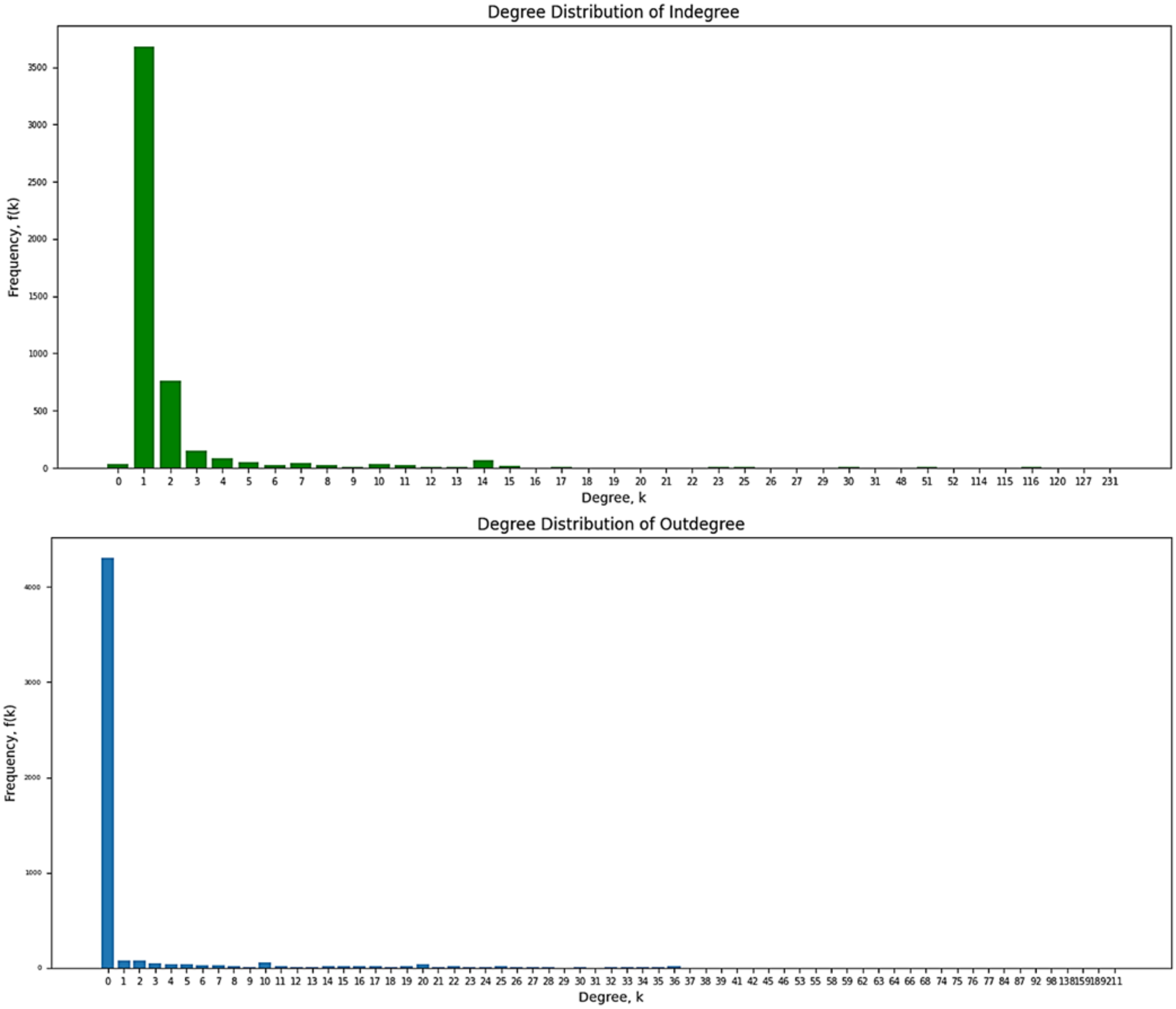

Fig. 1. Distribution of degree frequency, f(k), for both inwards (indegree) and outwards (outdegree) edges.

TABLE I
Degree of Separation

Criteria	Value
Maximum Degree of Separation	13
Minimum Degree of Separation	1
Median of Degree of Separation Values	5
Mode of Degree of Separation Values	4
Average of Degree of Separation Values	5.0

1) Degree Centrality: Degree centrality is defined as the number of links incident upon a node. It is the simplest centrality metric and is calculated locally. In case of directed graphs, there are tow type of degree for each node: indegree and outdegree. In this case, the degree centrality calculation might vary based on the network type. For web graphs, the degree centrality is classified into two cases: In degree Centrality and Out Degree Centrality [14]. In general, indegree is considered to be more important as it shows the dependency of other web pages on that particular web page. Here, Table II shows the top 10 nodes in the input network with the highest in degree and out degree centrality.

2) Closeness Centrality: Closeness Centrality represents how close a node is to other nodes in a network. It is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. In a network

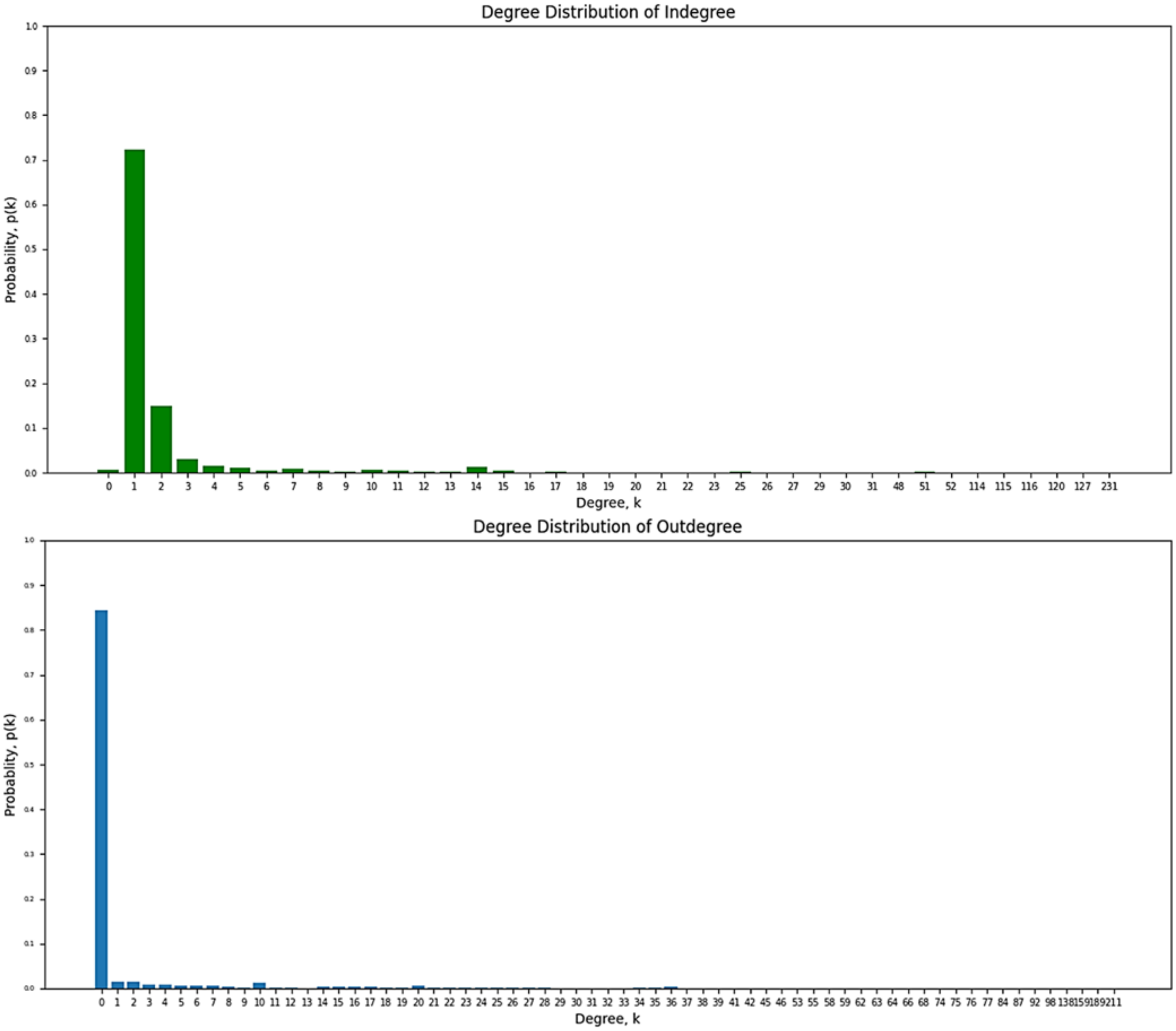

Fig. 2. Distribution of degree probability, p(k), for both inwards (indegree) and outwards (outdegree) edges.

TABLE II
TOP 10 NODES BASED ON IN DEGREE AND OUT DEGREE CENTRALITY

In Degree Centrality		Out Degree Centrality	
Node	Indegree	Node	Outdegree
226411	231	2656	211
234704	127	195916	211
105607	120	225872	189
38342	116	175144	159
81435	116	92583	138
167295	116	251990	98
198090	116	183865	92
214128	116	279100	92
34573	115	233947	87
245659	114	210403	84

containing V nodes, the closeness centrality of node s is calculated using Equation 1. Here, dist(s,t) represents the distance between the two nodes s and t. After normalizing closeness centrality using Equation 1, a high closeness centrality value indicates that the node is closely connected to other nodes in the network. The list of top 10 nodes based on highest closeness centrality value is shown in Table III.

$$C_{clo}(s) = \frac{1}{\sum_{t \in V} dist(s,t)} \tag{1}$$

3) Eccentricity Centrality: Eccentricity Centrality is very similar to closeness centrality. The only difference is that, instead of taking the summation of all the distances, it consideres the maximum distance of a node from the other nodes. The Equation 2 is used to calculate Eccentricity Centrality of node s in a graph containing V nodes. Table III shows the list of top 10 nodes based on highest eccentricity centrality value.

$$C_{ecc}(s) = \frac{1}{max\{dist(s,t) : t \in V\}} \tag{2}$$

TABLE III
TOP 10 NODES BASED ON CLOSENESS CENTRALITY, ECCENTRICITY CENTRALITY AND BETWEENNESS CENTRALITY

Closeness Centrality		Eccentricity Centrality		Betweenness Centrality	
Node	Centrality Value	Node	Centrality Value	Node	Centrality Value
6548	1.0	6548	1.0	226411	0.016
15409	1.0	15409	1.0	112742	0.008
194146	1.0	18412	1.0	17737	0.005
240934	1.0	149627	1.0	243109	0.004
55087	1.0	194146	1.0	91	0.003
256298	1.0	240934	1.0	232712	0.003
19960	1.0	55087	1.0	219782	0.003
204648	1.0	149993	1.0	204562	0.003
95363	1.0	13	1.0	145892	0.002
181503	1.0	41825	1.0	13719	0.002

4) Shortest Path Betweenness Centrality: Shortest path betweenness centrality, also known as betweenness centrality, quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. It captures how much a given node is in-between other nodes. The target node would have a high betweenness centrality if it appears in many shortest paths. Table III shows the list of top 10 nodes having highest shortest path betweenness centrality value compared to all other nodes for the input network.

C. Ranking Webpages Using Algorithm

1) PageRank Algorithm (PR): PageRank algorithm was introduced by one of Google's founders, Lary Page and is used to rank web pages in the Google search engine [15]. In order to get a rough estimate of importance of web-pages, PageRank counts the number and quality of links to a page. The underlying assumption is that more important websites are likely to receive more links from other websites. The equation used by the PageRank algorithm is given below:

$$PR(p_i) = \frac{1-d}{n} + d(\sum_{p_j \in V(p_i)} \frac{PR(p_j)}{C(p_j)}) \tag{3}$$

Here, PR(p_j) is the current PageRank value of page p_j, V(p_i) is the set of pages that have outward edges towards page p_i, C(p_j) is the number of outbound links from page p_j, n is the total number of pages and d is the damping factor ranging between 0 to 1. It required 29 iterations to rank the web pages of the input web graph using PageRank algorithm. The list of top 10 web pages with the maximum PageRank values have been listed in Table IV.

2) Weighted PageRank Algorithm (WPR): Weighted PageRank Algorithm is an extended version of the original PageRank algorithm. PageRank algorithm considers all links equal when distributing rank scores. But Weighted PageRank algorithm (WPR) takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages [9]. The equation used by the Weighted PageRank Algorithm is given below:

$$PR(p_i) = \frac{1-d}{n} + d(\sum_{p_j \in V(p_i)} \frac{PR(p_j)}{C(p_j)} * W^{in}_{p_j} * W^{out}_{p_j}) \tag{4}$$

Here, $W_{in}(p_j, p_i)$ is the weight of link (p_j, p_i) calculated based on the number of inlinks of page p_i and the number of inlinks of all reference pages of page p_j. Similarly, $W_{out}(p_j, p_i)$ is the weight of link (p_j, p_i) calculated based on the number

TABLE IV
THE LIST OF TOP 10 RANKED WEB PAGES USING PAGERANK, WEIGHTED PAGERANK AND HITS ALGORITHM

PageRank Algorithm		Weighted PageRank Algorithm		HITS Algorithm			
Node	PageRank Score	Node	Weighted PageRank Score	Node	Authority Score	Node	Hub Score
2	4.4×10^{-5}	2	1.5×10^{-5}	226411	0.05	7447	0.0098
226411	1.4×10^{-5}	226411	1.0×10^{-5}	234704	0.042	12569	0.0098
105607	6.1×10^{-6}	241454	5.31×10^{-6}	105607	0.042	17217	0.0098
38342	5.9×10^{-6}	89073	4.8×10^{-6}	81435	0.042	71996	0.0098
234704	5.7×10^{-6}	225872	4.1×10^{-6}	198090	0.041	93900	0.0098
167295	5.6×10^{-6}	105607	3.9×10^{-6}	214128	0.041	102533	0.0098
81435	5.2×10^{-6}	119479	2.8×10^{-6}	167295	0.041	162822	0.0098
198090	5.2×10^{-6}	124470	2.8×10^{-6}	34573	0.041	167050	0.0098
214128	5.2×10^{-6}	91620	2.2×10^{-6}	38342	0.041	179974	0.0098
34573	5.0×10^{-6}	192935	2.2×10^{-6}	245659	0.041	190554	0.0098

of outlinks of page p_i and the number of out-links of all reference pages of page p_j. According to a study [9], WPR performs better than the conventional PageRank algorithm in terms of returning a larger number of relevant pages to a given query. Table IV shows the list of top 10 web pages generated using the Weighted PageRank algorithm.

3) Hyperlink-Induced Topic Search Algorithm (HITS): Hyperlink-Induced Topic Search (HITS) algorithm, developed by Jon Kleinberg, measure the importance of pages or documents [10]. According to this algorithm, each web page has two scores: hub score and authority score. Hub-type pages are those that, although do not provide much information on a topic, link to the pages that do. Authority type pages are those that contribute content on a topic to a website and are therefore linked by many hubs pages related to that topic. In other words, pages containing useful information are considered to be authorities and pages that link to authorities are hubs. So, authority score represents quality of a page as content while hub score represents the quality of a page as an expert. At each iteration, hub score is calculated as the total sum of votes of authorities pointed to and authority score is calculated as the total sum of votes coming from experts and it continues until it reaches the convergence criteria. The list of top 10 web pages based on both authority score and hub score has been shown in Table IV.

D. Network Clustering

A cluster is a closely related group of objects. In order to identify the strongly related portions within a network, different clustering algorithms are used. Here, I have used the well known Markov Cluster Algorithm (MCL) [16] to identify the strongly related nodes. MCL is a fast and scalable unsupervised cluster algorithm for graphs. It uses Random Walk to find clusters because random walk is more likely to move around in the same cluster than to cross clusters. This is because, by definition clusters are internally dense while being separated by sparse regions. For this project, I have run the MCL algortihm using the python Networkx module and generated clusters within the input network. The 5094 nodes in the input network has been divided into 114 clusters. Figure 3 shows the histogram plot of distribution of nodes in clusters. Form the figure, it can be seen that the number of member nodes for majority share of clusters range between 1-20. The largest cluster contains 414 members nodes. The cluster division of all the nodes has been shown in Figure 4 where each cluster has been represented using a different color. From this figure, it can be clearly seen that there are nodes in the network which are forming internal closely connected regions. The nodes within an internal region are more densely connected to each other compared to the nodes outside the internal region.

E. Analyzing and Comparing the Results

From Table IV, it can be seen that the top 10 node list generated by the different algorithms are significantly different. However, there are some common nodes which have been assigned high score by majority of the algorithms. For example, node number ‘226411’ and ‘105607’ belong to the top-10 list of all the three algorithms. Node ‘2’ is common between PR and WPR algorithm. An interesting fact is that, the top-10 list based on PR algorithm and authority score of HITS algorithm are almost same. Among the top 10 nodes, 9 of the nodes are common based on PageRank Score and Authority Score. The top 10 node list generated using hub score of HITS algorithm is completely different from the top 10 lists generated by other scores. Also, Table VI, VII, VIII, IX show the degree centrality, closeness centrality, eccentricity centrality and betwenness centrality values for the top 10 nodes generated by the PR, WPR, HITS authority score and hub score respectively. Here, these centrality values have been considered as evaluating parameters for different top 10 web page lists generated by different algorithms. Based on the centrality values, it can bee seen that PageRank algorithm performs best as it has comparatively higher centrality values. The performance of HITS algorithm ranking based on authority score is almost as good as PageRank algorithm. Finally, Table V shows the number of iterations required for each algorithm in order to reach convergence. The number of iterations required by WPR algorithm is significantly less compared to other two algorithms. But given the same

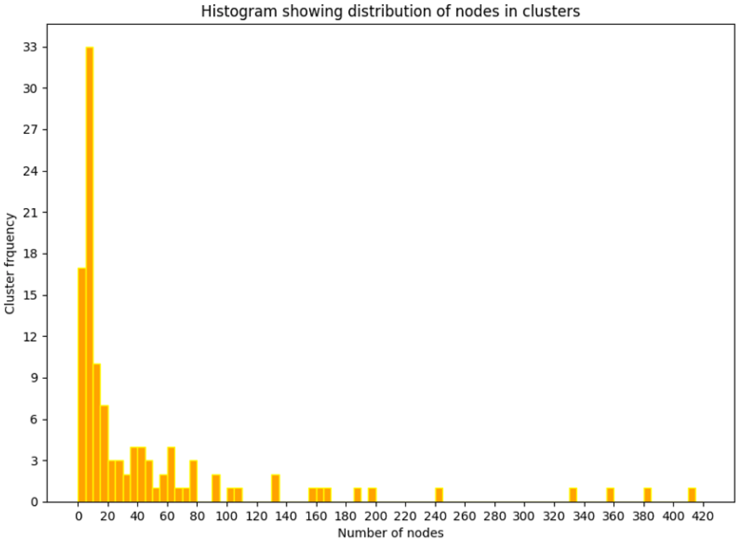

Fig. 3. Distribution of nodes among the clusters using MCL clustering algorithm

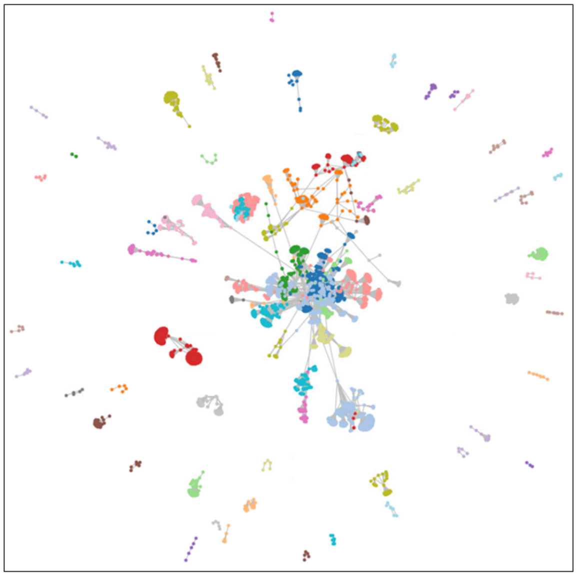

Fig. 4. Partition of the network nodes into clusters using MCL clustering algorithm. Here, different clusters have been represented using different colors.

environment, the computational runtime of WPR algorithm was significantly large compared to the other two algorithms and HITS was the fastest among these three algorithms.

TABLE V
COMPARISON OF ALGORITHMS BASED ON NUMBER OF ITERATIONS

Algorithm	Number of Iterations
PageRank Algorithm	29
Weighted PageRank Algorithm	8
HITS Algorithm	16

TABLE VI
NETWORK CENTRALITY MEASURES OF THE TOP 10 WEB PAGES RANKED BY PAGERANK ALGORITHM

Node	PageRank Score	In Degree Centrality	Closeness centrality	Eccentric centrality	Betweeness centrality
2	4.4×10^{-5}	31	0.0256	0.5	4.05×10^{-5}
226411	1.4×10^{-5}	231	0.0002	0.125	0.0157
105607	6.1×10^{-6}	120	0.0123	0.5	0.0007
38342	5.9×10^{-6}	116	0.0002	0.111	0.0002
234704	5.7×10^{-6}	127	0.0001	0.111	0.0006
167295	5.6×10^{-6}	116	0.0001	0.1	4.92×10^{-5}
81435	5.2×10^{-6}	116	0.0085	0.333	9.71×10^{-5}
198090	5.2×10^{-6}	116	0.0093	0.333	0.0002
214128	5.2×10^{-6}	116	0.0085	0.333	8.54×10^{-5}
34573	5.0×10^{-6}	115	0.0001	0.1	2.24×10^{-5}

VI. DISCUSSION

Network Centrality analysis for Web graph can be helpful in identifying important web pages in different application fields such as Product marketing, Commercial Advertisement, Recommender systems. The ranking algorithms are vastly used by web search engines for identifying web pages based on relevance and popularity. Clustering algorithms can be used to cluster the nodes so that each cluster can represent a group of similar web pages. Here, I have used all these above mentioned techniques to analyze the input web graph network. One weakness of this work is that, due to computational limitation, only a small portion of the whole network have been analyzed. Analyzing the whole network might have given some more useful insight

TABLE VII
Network Centrality Measures of the top 10 web pages ranked by Weighted PageRank algorithm

Node	Weighted PageRank Score	In Degree Centrality	Closeness centrality	Eccentric centrality	Betweeness centrality
2	1.5×10^{-5}	31	0.0256	0.5	4.05×10^{-5}
226411	1.0×10^{-5}	231	0.0002	0.125	0.0157
241454	5.31×10^{-6}	21	0.0097	0.333	6.34×10^{-5}
89073	4.8×10^{-6}	16	0.0065	0.5	0.0001
225872	4.1×10^{-6}	22	0.0001	0.1	0.0008
105607	3.9×10^{-6}	120	0.0123	0.5	0.0007
119479	2.8×10^{-6}	9	0.0357	1.0	9.72×10^{-6}
124470	2.8×10^{-6}	7	0.0001	0.111	0.0003
91620	2.2×10^{-6}	4	0.25	0.5	1.54×10^{-7}
192935	2.2×10^{-6}	4	0.25	0.5	1.54×10^{-7}

TABLE VIII
Network Centrality Measures of the top 10 web pages ranked by HITS algorithm's authority score

Node	PageRank Score	In Degree Centrality	Closeness centrality	Eccentric centrality	Betweeness centrality
226411	0.05	231	0.0002	0.125	0.0157
234704	0.042	127	0.0001	0.1111	0.0006
105607	0.042	120	0.0123	0.5	0.0007
81435	0.042	116	0.0085	0.3333	9.71×10^{-5}
198090	0.041	116	0.00993	0.3333	0.0002
214128	0.041	116	0.0085	0.3333	8.54×10^{-5}
167295	0.041	116	0.0001	0.1	4.92×10^{-5}
34573	0.041	115	0.0001	0.1	2.24×10^{-5}
38342	0.041	116	0.0001	0.1111	0.0002
245659	0.041	114	0.0001	0.1111	2.61×10^{-5}

TABLE IX
Network Centrality Measures of the top 10 web pages ranked by HITS algorithm's hub score

Node	PageRank Score	In Degree Centrality	Closeness centrality	Eccentric centrality	Betweeness centrality
7447	0.0098	2	0.0001	0.1111	1.43×10^{-7}
12569	0.0098	2	0.0001	0.1111	1.43×10^{-7}
17217	0.0098	2	0.0001	0.1111	1.43×10^{-7}
71996	0.0098	2	0.0001	0.1111	1.43×10^{-7}
93900	0.0098	2	0.0001	0.1111	1.43×10^{-7}
102533	0.0098	2	0.0001	0.1111	1.43×10^{-7}
162822	0.0098	2	0.0001	0.1111	1.43×10^{-7}
167050	0.0098	2	0.0001	0.1111	1.43×10^{-7}
179974	0.0098	2	0.0001	0.1111	1.43×10^{-7}
190554	0.0098	2	0.0001	0.1111	1.43×10^{-7}

about the dataset. It was seen, that the important nodes are not always same based on different centrality measures as different measures select important nodes based on different criteria. Also, the output of different ranking algorithms vary a lot. This points to the fact that different ranking algorithms might work better based on different requirement and usage purpose. In summary, different analysis techniques might come more handy in different scenarios and environment.

References

[1] K. Farrall, "Web graph analysis in perspective: Description and evaluation in terms of krippendorff's conceptual framework for content analysis (version 1.0)," *Retrieved on*, vol. 7, 2006.

[2] M. Shinde and S. Girase, "A survey of various web page ranking algorithms," *International Journal of Computer Applications*, vol. 975, p. 8887, 2015.

[3] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal, "The web as a graph," in *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2000, pp. 1–10.

[4] "How many websites are there?" https://websitesetup.org/news/how-many-websites-are-there/, accessed: 2021-12-4.

[5] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[6] K. Bharat, "Mr henzinger. improved algorithms for topic distillation in hyperlinked environments," in *Proc 21st Annual Intl ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 104–111.

[7] S. Yang, "Networks: An introduction by mej newman: Oxford, uk: Oxford university press. 720 pp., $85.00." 2013.

[8] D. Sullivan, "What is google pagerank? a guide for searchers & webmasters," *Search engine land*, 2007.

[9] W. Xing and A. Ghorbani, "Weighted pagerank algorithm," in *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.* IEEE, 2004, pp. 305–314.

[10] J. M. Kleinberg *et al.*, "Authoritative sources in a hyperlinked environment." in *SODA*, vol. 98. Citeseer, 1998, pp. 668–677.

[11] J. Leskovec and R. Sosič, "Snap: A general-purpose network analysis and graph-mining library," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 1, p. 1, 2016.

[12] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[13] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[14] B. Jaganathan and K. Desikan, "Enhanced web page ranking method using laplacian centrality," *International Journal of Engineering & Technology*, vol. 7, no. 4.10, pp. 566–569, 2018.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[16] S. vanDongen, "A cluster algorithm for graphs," *Information Systems [INS]*, no. R 0010, 2000.