APPLICATION OF GRAPH THEORETIC MEASURES IN REAL WORLD NETWORKS

BY
Nitish Kumar Sharma (Enrolment No. 19/11/EC/012)
Jha Aditya Dilip (Enrolment No. 19/11/EC/018)

under the guidance of
Prof. Krishnan Rajkumar, School of Engineering JNU, Delhi

in the partial fulfilment of the requirements
for the award of the degree of

Bachelor of Technology
(a part of Five-Year Dual Degree Course)

School of Engineering
Jawaharlal Nehru University, Delhi
Feb, 2023

JAWAHARLAL NEHRU UNIVERSITY

SCHOOL OF ENGINEERING

DECLARATION

We declare that the project work entitled **"APPLICATION OF GRAPH THEORETIC MEASURES IN REAL WORLD NETWORKS"** which is submitted by us in partial fulfillment of the requirement for the award of degree B.Tech. (a part of Dual-Degree Programme) to School of Engineering, Jawaharlal Nehru University, Delhi comprises only our original work and due acknowledgement has been made in the text to all other material used.

(Full Name and Sign of all Group Members)

JAWAHARLAL NEHRU UNIVERSITY

SCHOOL OF ENGINEERING

CERTIFICATE

This is to certify that the project work entitled **"APPLICATION OF GRAPH THEORETIC MEASURES IN REAL WORLD NETWORKS"** being submitted by **Mr. Nitish Kumar Sharma**(Enrolment No.- 19/11/EC/012) **& Mr. Jha Aditya Dilip**(Enrolment No.- 19/11/EC/018) in fulfilment of the requirements for the award of the **Bachelor of Technology** (part of Five-Year Dual Degree Course) in **Computer Science Engineering**, will be carried out by them under my supervision.

In my opinion, this work fulfills all the requirements of an Engineering Degree in respective stream as per the regulations of the School of Engineering, Jawaharlal Nehru University, Delhi. This thesis does not contain any work, which has been previously submitted for the award of any other degree.

Dr. Krishnan Rajkumar
(Supervisor)
Assistant Professor
School of Engineering
Jawaharlal Nehru University, Delhi

JAWAHARLAL NEHRU UNIVERSITY

SCHOOL OF ENGINEERING

ACKNOWLEDGMENT

We would like to express our gratitude towards Professor Krishnan Rajkumar, School of Engineering, Jawaharlal Nehru University, Delhi, for his invaluable guidance, support and the independence to pursue our curiosities.

At this juncture, when our project nears completion, we would also like to acknowledge the indomitable spirit that resides in all of us and holds us straight in times when we feel like falling inexorably. May this spirit inhabit the earth as long as there are curious beings striving to explore the recesses of knowledge.

"To strive, to seek, to find, and not to yield."

~ Alfred, Lord Tennyson

Nitish Kumar Sharma (Enrolment No. 19/11/EC/012)

Jha Aditya Dilip (Enrolment No. 19/11/EC/018)

LIST OF CONTENTS

Content
Page No.

LIST OF FIGURES

Details of Figure **Page No.**

ABSTRACT

Centrality in a network is the measure of how important a node or a vertex is in the network. There are different ways to define importance and subsequently there are different types of centrality based on which factor one deems important in a network. Modularity is a measure of the strength of connections within clusters contained in a graph i.e a graph of high modularity will have densely connected communities of nodes that will have sparse connections to other nodes in the graph. We study and review these two graph theoretic metrics and use them to analyse the network amongst user communities(called subreddits), obtained through the use of the python API wrapper of reddit(PRAW), on a social media platform called Reddit. The analysis consists of finding how influential a subreddit is using centrality measures and to find groups or communities of subreddits that are similar in their sets of users using the measure of modularity.

CHAPTER-1

INTRODUCTION and **THESIS OVERVIEW**

1.1. INTRODUCTION

Graph Theory is the branch of mathematics that studies graphs(also called networks), which are a set of vertices, also known as nodes, connected to each other through a set of edges. Graphs can be used to model many real life networks from social and informational to physical and biological. There are certain graph theoretic metrics, such as centrality and modularity, which aid us in understanding networks[1].

In a network, there are some nodes that are more important or yield more relative influence than other nodes. Centrality is the measure of the importance of a particular vertex or vertices in a network and depending upon the various definitions of importance one can derive various centrality measures. To give an example, perhaps the simplest centrality measure is *degree centrality* which defines the importance of a node on the basis of the number of edges it has. So in Fig 1.1, the degree centrality of nodes 4,5 and 2 is 3, of nodes 1 and 3 is 2, of node 6 is 1. In the course of this paper, we use pagerank, betweenness centrality and current flow betweenness centrality based on the requirements of the network[1].

When we study real life networks, we often come across some set of nodes that are connected densely to each other but sparsely to other nodes in the network. Modularity is the measure of how strong the connections between these sets of nodes(called modules or communities) are in the network. Networks that have high modularity also contain dense connections within modules or communities but very sparse connections to other nodes in the network while networks with low modularity would be akin to random connections between nodes[2].

Reddit is a social news and discussion website where registered users can post and engage in discussions on a wide range of topics. The website is organized into various communities, known as subreddits, each of which is centred on a particular subject, such as news, entertainment, science, technology, sports, politics, and more.[3].

Users can submit content to a subreddit, including links to articles, images, videos, and text-based posts. Other users can then vote on these posts and comments, with the most popular content rising to the top of the subreddit.In addition to posting and voting on content, users can also comment on posts, engage in discussions with other users, and interact with moderators who oversee the community. Reddit is known for its active user base and diverse range of content, making it a popular destination for people looking for information, entertainment, or discussion on virtually any topic.

We use graph theoretic metrics such as centrality and modularity to study the network of subreddits on reddit and identify the subreddits which possess the potential to make content viral on reddit.

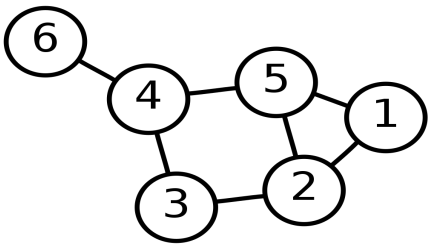

Fig 1.1

1.2. THESIS OBJECTIVE

The primary objective of this thesis is to

(i) gain insights into the networks of subreddits on reddit

(ii) help identify influential, in the context of their potential to make content viral, subreddits

1.3. ORGANIZATIONS OF CHAPTERS

The thesis consists of eight chapters and the overview of all the chapter are as follows:

Chapter 1: This chapter provides a brief introduction on the background, the objectives of the thesis involved in accomplishing the thesis.
Chapter 2: This chapter gives an overview of available literature.
Chapter 3 describes the details of proposed work and methodology used in this research work.
Chapter 4: This chapter gives the discussion about the simulated and measured results.
Chapter 5 presents the conclusion of this research work and future scope of the work.
References: In the last section of this thesis, used references are given.

CHAPTER-2
LITERATURE SURVEY

2.1. INTRODUCTION

A graph G, known as a network, is comprised of two separate sets (V,E) that are ordered and do not overlap. V is a collection of elements referred to as nodes or vertices, while E is a subset of ordered pairs of distinct elements from V that are identified as edges. Additionally, the edges in the network may contain weights that reflect the intensity of the association between the two nodes.We encounter networks in various settings in the real world with most of the networks exhibiting very heterogeneous structure and thus making the identification of important nodes crucial. Most real life networks also exhibit the presence of different communities within them.

2.2 Centrality

In complex real world networks, due to it possesing a highly heterogeneous structure as mentioned above, some nodes can be regarded as being more important in the network than the other nodes. Due to this their identification becomes a crucial task for understanding any network.For instance, in a network of university students, some people, like the Vice-Chancellor or Rector and student leaders yield a lot of influence amongst the students as compared to others. As a consequence, these nodes can be regarded as being central in the network.

However, the definition of centrality is not absolute, as we potentially can define it in many ways depending on what we are seeking to find in the network. For example, in the above case of the network of university students, the Vice-Chancellor yields influence because of her formal powers that accompany her position while the student leader becomes influential due to the sheer number of interactions that she has with different students in the university. Thus, the definition of centrality cannot be universal and must depend on the particular real-life application. But, since we do not have an agreement about the general definition of centrality, several definitions have been proposed, based on what one considers to be 'important' vis-à-vis the network[4].

2.2.1. PageRank

Pagerank centrality is a network centrality measure developed by Google co-founders Larry Page and Sergey Brin for ranking web pages on the internet. Pagerank centrality assigns a score to each node in a network based on the number and quality of links pointing to it. The more edges a node has coming into it and the more important the nodes coming into it, the higher is its Pagerank score[5].

Let G = (V, E) be a directed graph with n nodes and m edges, where V is the set of nodes and E is the set of directed edges. Let A be the adjacency matrix of G, where A[i, j] = 1 if there is a directed edge from node i to node j, and 0 otherwise.

The Pagerank score of node i, denoted by PR(i), is defined as:

$$PR(i) = \frac{(1-d)}{n} + d \sum_{j \in B_i} \frac{PR(j)}{L(j)}$$

where d is the damping factor, typically set to 0.85, which represents the probability that a user will continue to browse the web by following a link, and $\frac{(1-d)}{n}$ is the probability that the user might jump to a random page. L(j) is the number of outgoing links from node j, and B_i is the set of nodes that link to node i.

Interpreting PageRank values can provide insights into the structure and behavior of complex networks, as well as the potential impact of node removal. Nodes with high PageRank values act as important hubs within the network, often connecting separate parts of the network and facilitating the flow of information or resources between nodes. These nodes are also more likely to have a higher impact on the network's overall behavior and function, making them potential targets for intervention or disruption[1].

Pagerank centrality has been used in various studies to analyze networks and identify important nodes. For example, a study by Newman et al. (2001)[6] used Pagerank centrality to analyze the structure of the World Wide Web and found that a small number of highly ranked nodes (known as "hubs") play a crucial role in connecting different parts of the web. Another study by Wu et al. (2015)[7] used Pagerank centrality to identify important nodes in a social network and found that these nodes are more likely to be influential in spreading information or opinions.

2.2.2. Betweenness

Betweenness centrality is an another measure of a node's value or importance in a network based on how often it exists on the shortest paths between pairs of other nodes[8]. A node with a high value of betweenness centrality lies on many shortest paths and is thus important for the communication and flow of information in the network.

The betweenness centrality of a node i is defined as:

$$C_B(i) = \sum_{s \neq i \neq t \in V} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

where V is the set of nodes in the network, σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st}(i)$ is the number of those paths that pass through node i. The centrality

of betweenness of a node is the sum of the fraction of all pairs of nodes that pass through it.

The computation of betweenness centrality can be computationally expensive for large networks, especially when using the exact calculation formula. Therefore, various approximate algorithms have been developed, such as Brandes' algorithm[9], which gives a time complexity of O(nm), where n is the number of nodes and m is the number of edges in the network.

Betweenness centrality has been applied in various fields, such as the analysis of social networks, transportation planning, and protein interaction networks. In social networks, nodes with high betweenness centrality are very crucial for their potential for the spread of information or disease. In transportation planning, nodes with high betweenness centrality are crucial for efficient transportation and congestion management. In protein interaction networks, nodes with high betweenness centrality are potential drug targets for treating diseases[1].

2.2.3. Current Flow Betweenness

Current flow betweenness centrality is a measure of a node's importance in a network based on the flow of electrical current between pairs of other nodes[9]. A node with high current flow betweenness centrality plays a crucial role in the transmission of current and information in the network.

The current flow betweenness centrality of a node i is defined as:

$$C_{CF}(i) = \sum_{s \neq i \neq t \in V} \frac{\Phi_{st}(i)}{V_{st}}$$

where V is the set of nodes in the network, $\Phi_{st}(i)$ is the current flow passing through node i from node s to node t, and V_{st} is the voltage between nodes s and t.

Current flow betweenness centrality has been applied in various fields, such as social network analysis, power grid analysis, and protein interaction networks. In social networks, nodes with high current flow betweenness centrality are important for the spread of information or opinion. In power grid analysis, nodes with high current flow betweenness

centrality are critical for the transmission of electrical power. In protein interaction networks, nodes with high current flow betweenness centrality are potential drug targets for treating diseases[10].

2.3 Modularity

The measure of modularity assesses the degree to which a network can be segregated into smaller subgroups, or communities, based on the pattern of connections between nodes. Modularity is commonly used in community detection, which is the process of recognizing densely connected subgroups of nodes within a network[11].

The modularity of a network with n nodes and m edges can be computed using the following formula:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$$

where A_{ij} is the element of the adjacency matrix that indicates whether nodes i and j are connected, k_i and k_j are the degrees of nodes i and j respectively, m is the total number of edges in the network, c_i and c_j are the community assignments of nodes i and j respectively, and $\delta(c_i, c_j)$ is the Kronecker delta, which equals 1 if $c_i = c_j$ and 0 otherwise.

The modularity score ranges from -1 to 1, with higher values indicating a greater degree of community structure in the network. A modularity score of 0 indicates that the network has no detectable community structure[12]. The modularity value for graphs that are unweighted and undirected will lie, given the definition, in the range $[-1/2,1]$[13] .

Modularity has been applied in many fields, such as the analysis of social network and social media network, biology, and computer science. In social network analysis, modularity is used to identify groups of people with similar characteristics or interests. In biology, modularity is used to identify functional modules in biological networks such as gene regulatory networks or protein interaction networks. In computer science, modularity is used in software engineering to identify modules or components in software systems[14].

We use modularity in the context of community finding.Community finding is a process that consists of identifying groups of nodes in a network that are densely connected to each other but less connected to nodes outside the group. Modularity compares the actual number of edges within communities to the expected number of edges if the edges were to be say, hypothetically, distributed at random, with higher values indicating a stronger division of the network into communities.

CHAPTER-3

PROPOSED WORK AND METHODOLOGY

3.1. INTRODUCTION

We begin by using PRAW(The Python Reddit API Wrapper) to scrape data from reddit[15]. We get the top 1000 posts from various popular subreddits(for example r/india see fig 3.1), these act as the base subreddit from which all connections are derived, and extract information about their author using which we get the author's top 10 posts. Here, the top posts signify that the content is fairly popular and as a consequence, in most cases, their author's must have posted a lot of content before since getting a given post in top must be significantly subject to chance.

We clean the data and construct from it a graph using the NetworkX package in python where the nodes consist of individual subreddits and there is an edge between any two subreddit if a common user posts in them. All self edges are removed. This gives us a network of subreddits which are connected by their common posters. SInce, the author is one among the many users that have joined the community, it can be assumed that his/her choice to join and post in other subreddits may reflect the actions of significant numbers of users given that the author has posted popular content. Hence, the connection between the two subreddits can be interpreted as them having significant overlap of users.

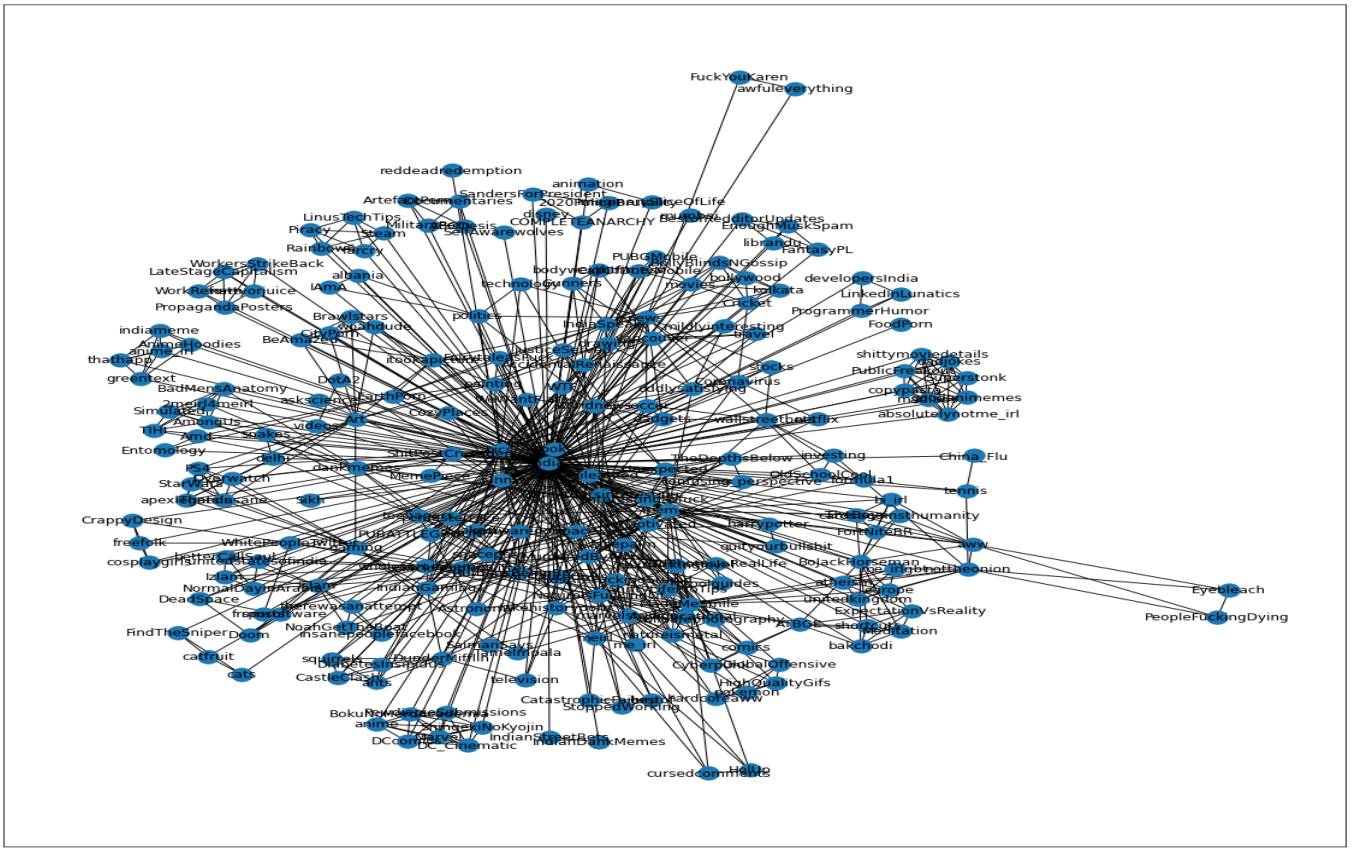

Fig 3.1 Network of subreddits from the base subreddit of r/india

On this network, we apply centrality measures as discussed in the previous section and draw the graph with labels where the size of each node denotes its centrality value. SInce, the base subreddit would naturally have the highest value of centrality and thus skew the data, it is scaled down so as to reduce its size in the representation.

The choice of the particular centrality measure was informed by [15] and eigenvector centrality was replaced by pagerank as its additional improvements seemed like a better measure. The inclusion of current flow betweenness centrality is because it is able to model flow of information(in this case content) through non-shortest paths and hence, gives us an important clue into the network.

Further, by using the measure of modularity we try to find communities in the network using the Clauset-Newman-Moore greedy modularity maximization algorithm[11] . We draw different communities in different colours. We then discuss the meaning of these communities and the value of modularity in the context of the network of subreddits.

3.2. Results

The output of the centrality measures discussed above on r/india is as follows:

3.2.1. Pagerank

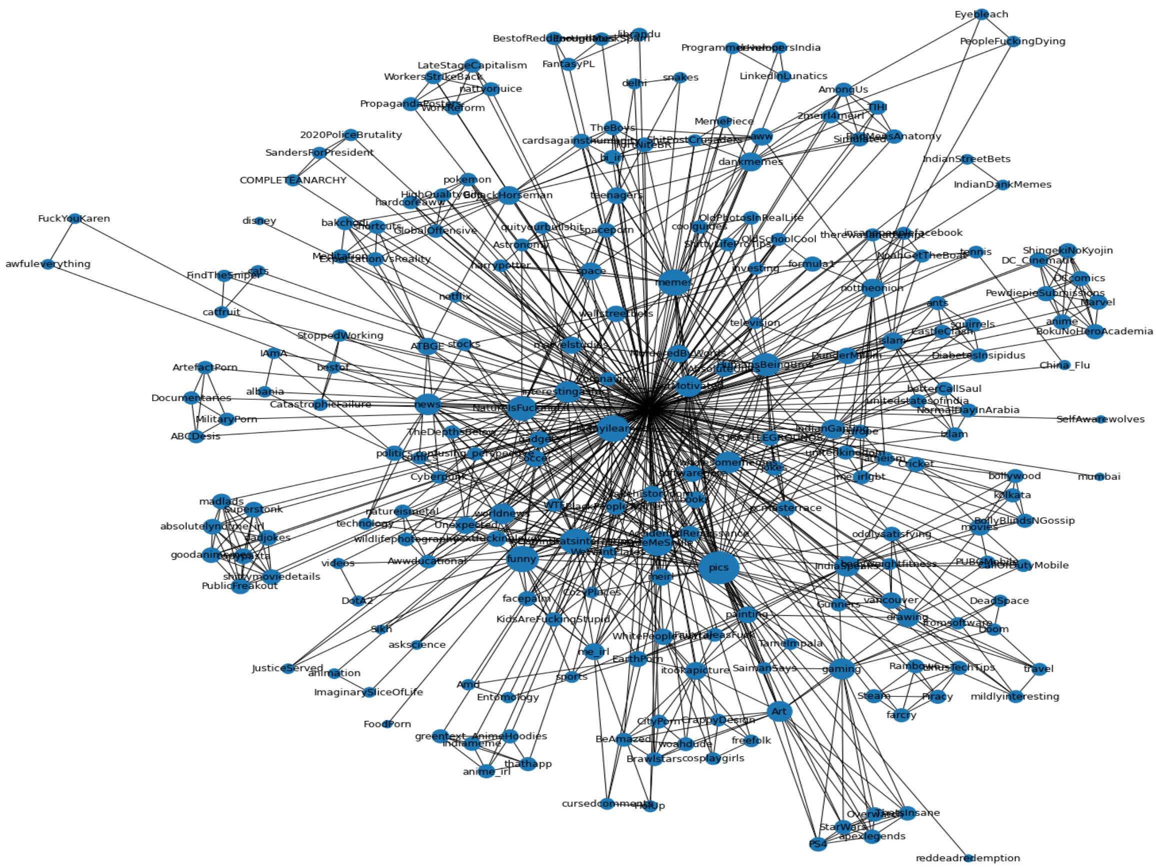

Fig 3.2

3.2.2. Betweenness

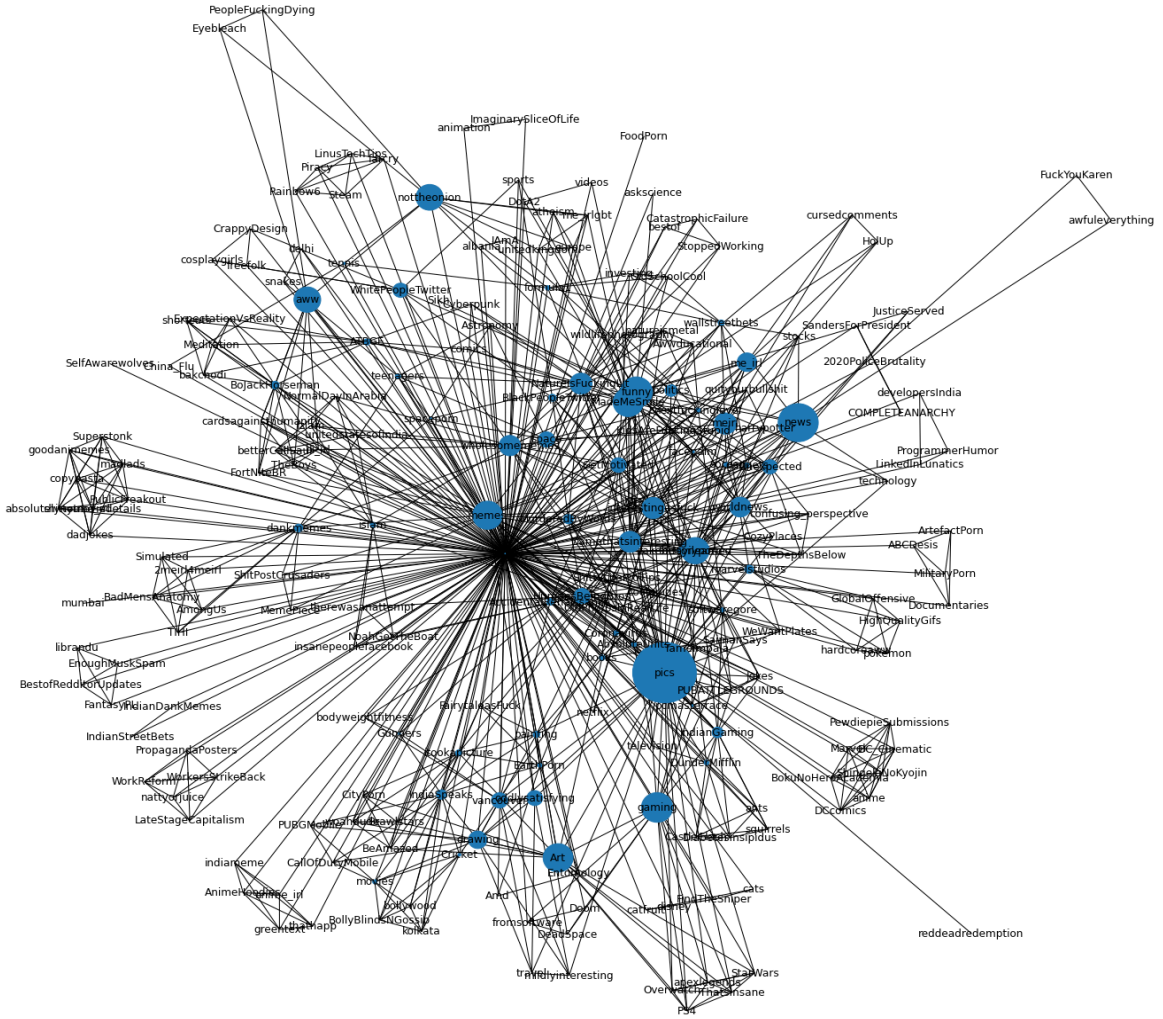

Fig 3.3

3.2.3. Current Flow Betweenness

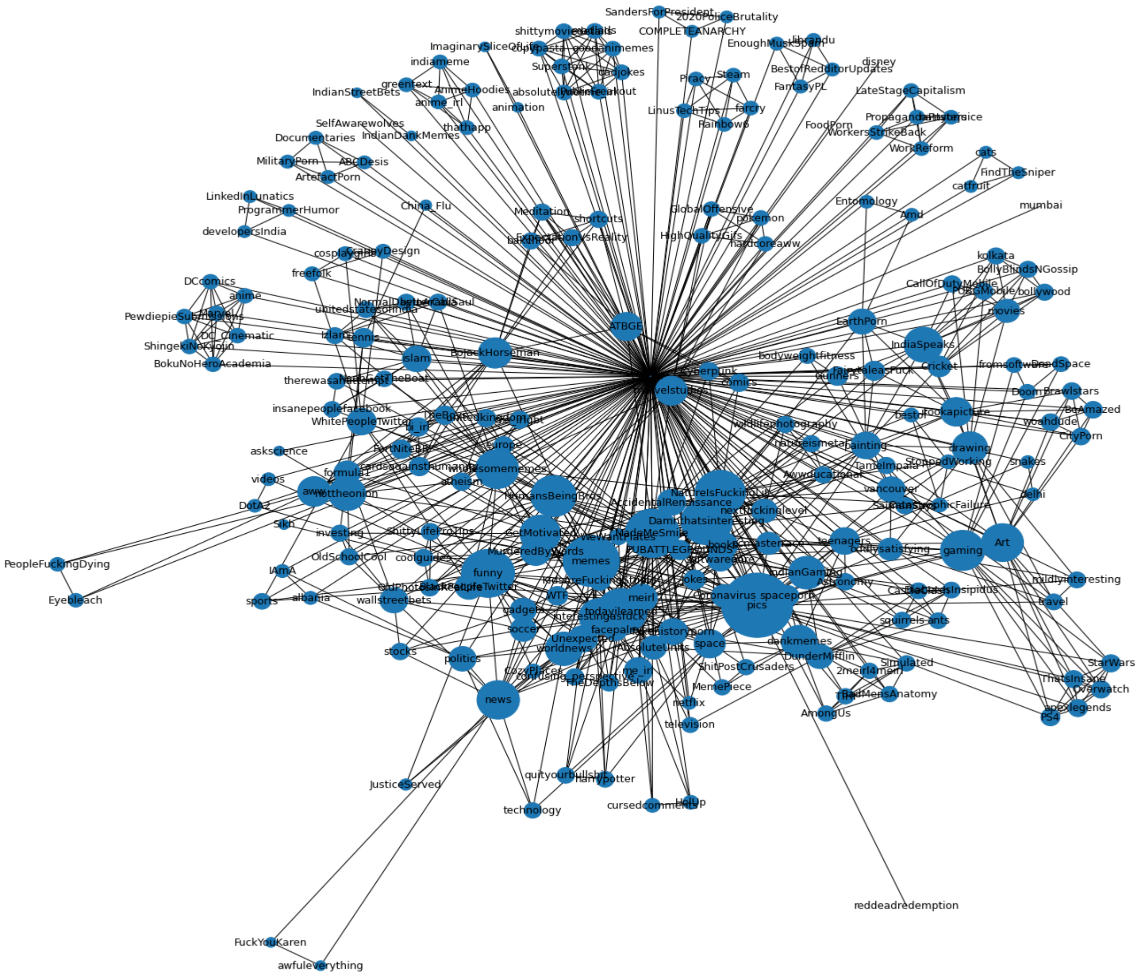

Fig 3.4

3.2.4. Communities

The following communities(12) were found on r/india

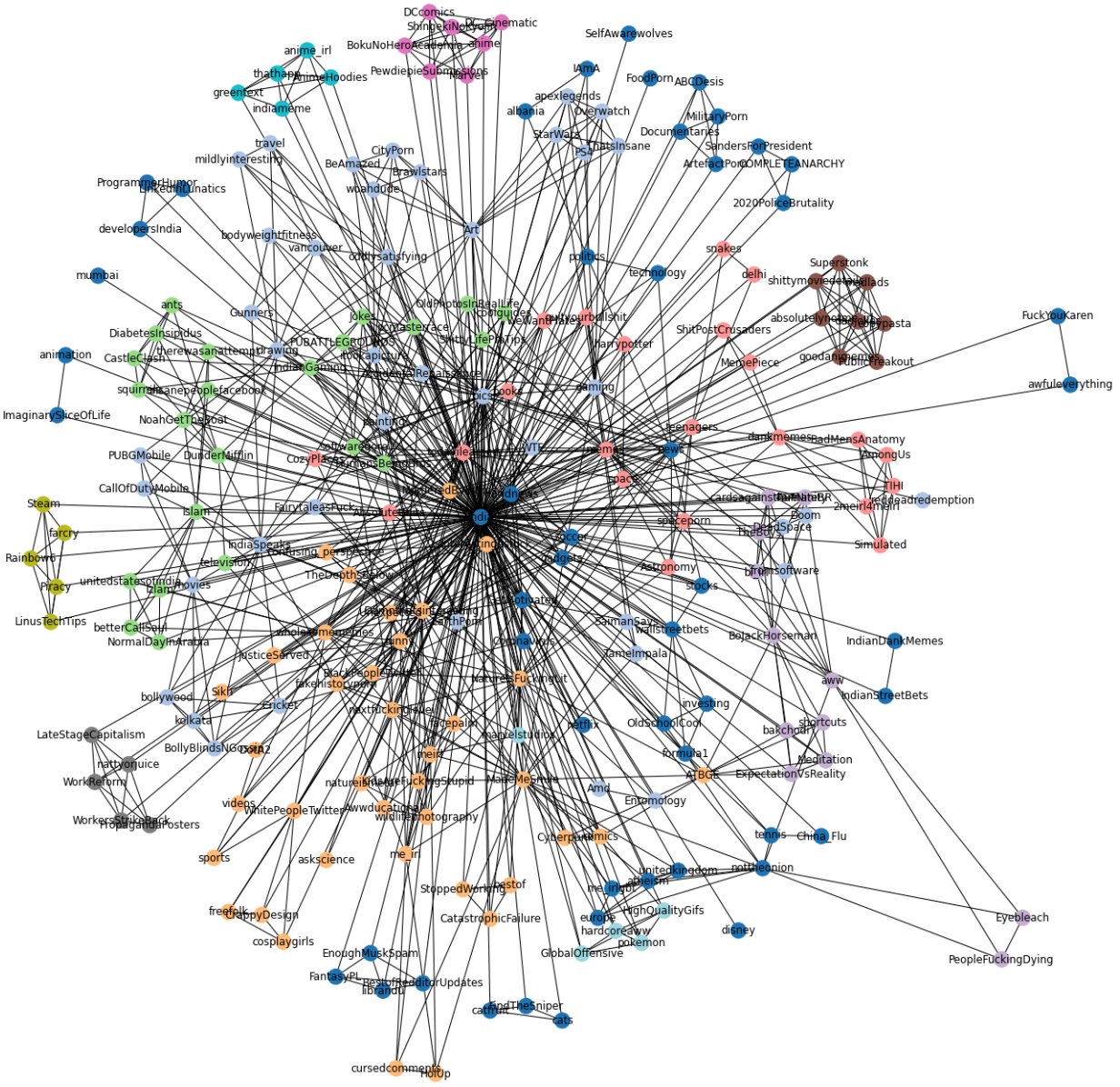

Fig 3.5

CHAPTER-4
RESULT DISCUSSION

4.1. INFLUENTIAL SUBREDDITS

Imagine a user **u** on reddit joining an initial subreddit **s,** he looks at the top posts of the subreddit to acquaint himself with its content(here defined as any media or text that can be

posted on reddit) . Suppose he clicks at the profile of an author **a,** who is the author of a top post, he will come across all the content the author has posted in the same as well as different subreddit. It is at this juncture that **u** will discover other subreddit and subsequently other content and will repeat the process described above recursively till he stops.

Here, the thing to note is that **u** starting from **s** visits a series of subreddits and through this process consumes contents of different kinds via different authors starting from **a.** As content is created to be consumed and its success is gauged by its consumers, we face the following question, “Which subreddit to post in so that on an average our content gets the most exposure?”.

To answer this question, we must first look into the following question, “What happens to most of the content that is posted on reddit?”. The answer from experience seems to be that most of it dies down in the subreddit itself while some of the content manages to be moderately popular and is also consumed by “like” subreddits i.e the subreddits that have a mostly similar set of users and form a community, in the context of the networks of subreddits.

Now, we define the influence of a subreddit to be its capacity to expose content to subreddits which have a dissimilar set of users than its own. The question asked above can then be rephrased to, “Which subreddits have the most amount of influence in the network?” and the answer to which can be attempted by network analysis.

4.2. PageRank

We observe in Fig 3.1 that pagerank values do not vary significantly across the network. This implies that no subreddit stands out on the basis of its connection to important nodes and that any subreddit in the network can be reached from any other subreddit. Subreddits in the network contribute relatively similarly towards the connectivity in the network.

If we assume that the user takes a random walk through the network, most of the nodes could be reached with a decent probability. Finally, this implies that content does circulate quite freely throughout the network. Posting in any given subreddit one has a non zero probability of it getting viral.

4.3. Betweenness

We observe in Fig 3.2 that few nodes have very significant values of betweenness centrality as compared to other nodes. The user **u** traversing the network through shortest paths would visit these nodes more frequently than others. This suggests that these nodes are the sites at which wide content exposure can take place and indicates that there exists nodes which are the gateway to virality.

These nodes with high value of betweenness centrality join different communities of subreddits in the network and hence are the places where very diverse sets of users come together. Any content posted in these subreddits may have a high rate of exposure and subsequently a high probability of getting viral and conversely they diminish the equal access of content to all users. Also, the minimal number of these subreddits suggests their importance.

The drawback of just considering only these subreddits would be our assumption of users exclusively taking shortest paths in the network which in reality, may not always be true. To address this we use the current flow betweenness, which by modelling flow as an electric current, also takes into consideration other paths between two subreddits rather than just taking the shortest path. This will allow us to get a more comprehensive list of subreddits that have influence in the network.

4.4. Current Flow Betweenness

We observe in Fig 3.3, along with the subreddits that featured prominently in betweenness centrality, we have other prominent subreddits with high values of centrality that were insignificant in Fig 3.2. Also, as compared to Fig 3.2 most nodes have significant value of centrality.

Subreddits with high current flow betweenness centrality values act as bottlenecks for the flow of content and may be critical for the network's overall performance. These subreddits can also act as potential failure points, as the network's robustness may be compromised if they are removed or attacked by malicious forces.

In the end, current flow betweenness centrality was found to be a valuable tool for understanding the structure and behavior of the network, as it provides insights into the role of subreddits in facilitating the flow of content through the network.Thus, current flow betweenness gives us a comprehensive list of subreddits that yield influence in the network because of their role as gatekeepers of content in the network.

4.5. Communities

In the context of our network, communities can be interpreted as groups of subreddits with similar sets of users and as a result content that can be homogeneous. These communities are important as if there are not enough linkages to outside nodes then these communities can trap content within them. This can result in them being echo-chambers where outside influence is quite rare. This can be undesirable especially in the age of fake news where people quite readily believe news that comes on their social media.

To know about communities of subreddits is also needed because in most cases of targeted content creation one doesn’t need to reach everywhere but only to a subsection of users that inhabit the virtual communities through these communities of subreddit.
If we are able to point out specific communities we can also deduce the kinds of users in the base subreddit. For example, users from a different country that are on r/india can be demonstrated by the other communities of subreddits joined by them.

Using the Clauset-Newman-Moore greedy modularity maximization algorithm we were able to split the network into 12 distinct communities as shown in Fig 3.4. The network had a relatively significant modularity of 0.49. This suggests that there exists communities within the network but not very strong ones.

We can deduce that the content is not bound or trapped in very rigid communities and that the network allows ample avenues for it to circulate and influence subreddits other than its own. This represents a desirable property in the network that it is not split up in modules and is relatively open globally.

CHAPTER-5

CONCLUSION AND FUTURE SCOPE

5.1. CONCLUSION

We can conclude the following about the network of subreddits on reddit:

a) It is relatively well connected and allows content to move around freely.
b) Content posted on any subreddit has a non zero probability of getting viral
c) In Spite of this, there exists subreddits that act as important gateways to content in the network
d) These subreddits are relatively few in number
e) Content posted on these subreddits have a higher probability of being viral
f) The network has relatively medium modularity
g) There exists communities of subreddits with similar sets of users but these communities are not immune from influences from other communities

5.2. FUTURE SCOPE

We can try to scale the dataset and capture the network in more detail which will allow us to carry out more minute analysis. We can also introduce new graph theoretic metrics that will further increase our understanding of these networks. Another avenue that could be traversed is the conditions for virality of content on other social media platforms.

REFERENCES

[1] https://math.bme.hu/~gabor/oktatas/SztoM/Newman_Networks.pdf
[2] https://en.wikipedia.org/wiki/Modularity_(networks)
[3] https://en.wikipedia.org/wiki/Reddit
[4] Rodrigues, Francisco. (2019). Network centrality: an introduction.
[5] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab
[6] Newman, M. E., Barabási, A. L., & Watts, D. J. (2006). The structure and dynamics of networks (Vol. 2). Princeton University Press.
[7] Wu, Y., Liu, J., Li, F., & Li, X. (2015). Identifying influential nodes in complex networks. Physica A: Statistical Mechanics and its Applications, 421, 514-524.
[8] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry, 40(1), 35-41.
[9] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 25(2), 163-177.
[10] Zhang, X., Wang, W., & Wang, J. (2015). Identifying critical nodes in complex networks with current-flow betweenness centrality. Scientific reports, 5(1), 1-10.
[11] Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2), 026113.
[12] Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3-5), 75-174.
[13]https://kops.uni-konstanz.de/server/api/core/bitstreams/283d523f-ebf3-4f1d-9bd0-96e4b96bc4eb/content
[14] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.
[15] S. P. Borgatti, "Centrality and network flow," Social Networks, vol. 27, no. 1, pp. 55-71, 2005/01// 2005.