VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

PROJECT REPORT

SOCIAL AND INFORMATION NETWORKS

CSE 3021

PROJECT TITLE –

SOCIAL NETWORK ANALYSIS OF FACEBOOK PAGES

TEAM MEMBERS-

1. Soumyaraj Roy (19BCE0200)
2. Kaustubh Dwivedi (19BCE0249)
3. Nishma Rebello (19BCE0253)
4. Ayush Pandya (19BCE2697)

ABSTRACT

Social networking sites like Facebook, Twitter, and Instagram are most visited domains on the Internet. They contain huge data about the users and the relationships among them. To analyse and mine useful information from these huge social network data, special graph-based mining tools are required that can easily model the structure of the social networks.

Several such analysis tools are available with their own features and benefits. Choosing an appropriate tool such as a programming language and its functions and libraries, we can use it to do social network analysis on Facebook pages.

We will use the R programming language for conducting the analysis on Facebook football related pages dataset and as our project deals with data analysis and statistics, R provides flexibility to use available libraries and hence suites our work.

INDEX PAGE

1. INTRODUCTION

In this project, we are going to use a graph dataset containing various data about number of Facebook pages of organizations directly or indirectly related to football. This information includes

a) fan_count (no. of likes of the page has received)

b) category (such as ‘Non-Profit Organization’, ‘Athlete’, ‘Company’, etc.)

c) username (assigned to distinguish every page)

d) users_can_post (can any Facebook user post to that page)

e) link (URL for the corresponding Facebook page)

f) post_activity

g) talking_about_count (how many people have been talking about this page recently)

In the proposed system, we will be generating graph statistics and be producing visualizations which will greatly aid the reader in understanding the subtle nuances of the things happening behind the scenes. We will gain a greater understanding of graph theory and social media analysis than we do now.

The useful insights that we can generate from such data, can be used by corporates and organizations to improve their business models, and do targeted marketing campaigns.

2. LITERATURE SURVEY

We have done a survey report of ten related work of research papers mainly from IEEE and noted down parameters such as Theoretical model, Methodology used, Relevant Finding and Limitations or Future Research.

Sno	Author and Year of Reference	Title	Concept / Theoretical model/ Framework	Methodology used/ Implementation	Relevant Finding	Limitations / Future Research/ Gaps identified
1	Kaing Sabai Phyu, Myat Myat Min. Published in: 2019 IEEE, 18th International Conference on Computer and Information Science (ICIS)	Graph-based Community Detection in Social Networks	The authors have provided a framework for sentiment community detection in twitter network. The main function of the system is to explore a way to find out sentiment communities which maximize both the intra-relations of vertices and the consistency of sentiment polarities within one community in twitter network.	The proposed system is made up of the five key phases: 1. Data Extraction via NodeXL. 2. Data Preparation using NLTK. 3. Sentiment Analysis. 4. Community Detection. 5. Sentiment Community Detection. They have calculated sentiment values of tweets by using three classifiers.	The Sentiment communities found through authors' methods can describe the various tastes of consumers, and vendors which can help to recommend similar products with positive views to communities and avoid recommending similar products to communities with opposing views.	The research is limited to twitter social networking site but can be further extended by applying the concept of sentiment community detection to other personal social networks like Facebook.

2	Nadiya Straton, Raghava Rao Mukkamala, Ravi Vatrapu. Published in: 2017 IEEE 6th International Congress on Big Data	Big Social Data Analytics for Public Health: Predicting Facebook Post Performance using Artificial Neural Networks and Deep Learning	The authors have tested different algorithms, models, and statistical approaches to find the most effective method to evaluate post performance and then make predictions based on the number of relevant attributes with Artificial Neural Networks and Deep Neural Networks.	The methodology is to predict the user engagement based on eleven characteristics of the post: Post Type, Hour Span, Facebook Wall Category, Level, Country etc. Finally, post-performance prediction was conducted using Artificial Neural Networks (ANN) and Deep Neural Networks (DNN).	1. The research study aims at health and care organizations to improve information dissemination on social media platforms by reducing clutter and noise. 2. At the same time, it will help users navigate through vast amount of information in direction of the relevant health and care content.	For future work, one would like to combine dynamic architecture with ANN and look into incremental clustering instead of traditional clustering approaches in order to be able to handle real-time data flow.
3	Hui Li. Published in: 2018 IEEE 3rd International Conference on Big Data Analysis	Centrality Analysis of Online Social Network Big Data	The authors discussed about Node centralities and network centralizations of online social networks .	The methodology followed is that of a typical online social network, where BBS reply network is firstly constructed and in which identifications registered on BBS are considered as nodes and reply articles are set up as links between them. Then, degree centrality and betweenness centrality of nodes are investigated, and the correlation between degree and betweenness centrality is mentioned.	1. Through the Measure of correlation coefficient of degree and betweenness centrality, it is found that nodes with high degree is inclined to have high betweenness centrality, thus betweenness centrality is positively related to degree. 2. Central nodes with high degree or high betweenness centrality do have high influence and power in online social networks	In this paper, static analyses of node attacks are only presented. In Future research, dynamical analyses of node attacks on online social networks can be focused upon.

4	Weichang Du Faculty of Computer Science, University of New Brunswick Fredericton, NB Canada	Toward Semantic Social Network Analysis For Business Big Data	The analysis on historical stock data identifies alternative representative indexing stock groups. The analysis on high frequency trading data establishes new algorithms for more effective high frequency trading. The analysis on business contract networks studies relationships between companies' contracts and their performance in profits and stock levels.	Analysis on historical stock data that identifies alternative representative indexing stock groups and possible stock change prediction. Describes the analysis on business contract networks that studies relationships between companies' contracts and their performance in profits and stock levels.	In this paper author have introduced three of our research projects which apply SNA techniques to business data analysis. We need to used relevant semantic models and establish relationships between social network and such semantic models.	The results of all the three research projects are positive, showing that SNA can be used as an effective technique on business related research and practices.
5.	Oksana Severiukhina, Klavdiya Bochenina National Center for Cognitive Research ITMO University Saint Petersburg, Russia	Segment-wise Users' Response Prediction based on Activity Traces in Online Social Networks.	The authors dealt with how behaviour of users in different activity segments varies and how predictable the behaviour of segments is. To analyse the reactions to posts, authors have used topic modelling and post tonality analysis. They have proposed a method for the response prediction to posts based on the prediction for separate activity segments and combining results	The methodology is to perform prediction of the aggregated dynamics of reactions in social systems on the example of user responses in thematic communities. The prediction method combines a history of users' actions in frames of a given online community with topical information of new messages to measure the expected reaction of the audience.	This method gives more accurate results than the general forecast for all subscribers in the community allows for each type of user to determine the most important characteristics of the post that affect the likelihood of a reaction. This allows us to determine the behavior of different segments of users depends on the activity.	For future research, one can go on determining the features and borders of segments for better prediction quality.

			prediction for separate activity segments and combining results.			

6.	D Radha, Shantanu Kulkarni Department of Computer Science and engineering Amrita School of Engineering, Bengaluru Amirta University India	A Social Network Analysis of World Cities and Network	The authors did an analysis related to the world cities network has been done. World city network presents a bipartite kind of graph. The parameters and methodologies for analysis are dependent on what type of data, a graph is representing. It also depends on the kind of a graph obtained from the network. World city network can be studied for analysis purpose.	Several graphical layouts available for better visualization of a network data. In several networks, best centrality measure can be decided by individual vertex centrality measure comparison. It depends on the type of nodes. Generally, a square adjacency matrix is expected for calculations. For nonsquare adjacency matrix, different 2-mode measures are available	This analysis is carried out with several centrality measures. Cities connected with more service firms, cities connected to the firms having attachments to larger cities, which firm is closest to all other cities, etc. are some details that can be obtained from the analysis.The network is analyzed by using different centrality measures to find which node is important or which node is more central and closer is implemented.	The network here discussed is a global information data. There are 315 cities, but the actual number may differ. This work is beneficial to understand the vital roles of the individual cities in particular aspects in the chosen network dataset.
7.	Rola Abdulkarim, Sherief Abdallah 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)	Using Social Network Analysis to Study Diversity in Business Partnerships	So in this paper we study the diversity in business partnerships as understanding it could be of significant economic value. And this would help a lot of policy makers to assess their policy and shape their policy accordingly..	First the 3 averages are found that are a)Average Degree b)Average Weighted Degree c)Average Edge Weight And after calculating these filtration of node takes place Nodes of degree higher than average Degree and edge weight higher than Average edge weight are selected.	While the business partnerships show diversity with respect to age and nationality, there is a lack of diversity when it comes to gender	applying the technique of link prediction to predict new or dissolved partnerships. Implementation of a link prediction can be extended to a recommendation system suggest a business partner to another.

				Communities are formed on the basis of betweeness centrality and Modularity.		
8.	Xiao Yang Search Data Technologies Seungbae Kim 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining	How Do Influencers Mention Brands in Social Media? Sponsorship Prediction of Instagram Posts	This paper produces a study of the brand mentioning practice of influencers. It analyzes a brand mentioning social network built on 18,523 Instagram influencers and 804,397 brand mentioning posts	They have used a neural network-based model to classify the sponsorship of posts utilizing network embedding and social media features. It gives a accuracy of 80%	(i)Most influencers mention only a few brands in their posts; (ii) popular influencers tend to mention only popular brands while micro-influencers do not have a preference on brand popularity; (iii) audience have highly similar reactions to sponsored and non-sponsored posts; and (iv) compared to non-sponsored posts, sponsored brand mentioning posts favor fewer usertags and more hashtags with longer captions to exclusively promote the specific products	They plan to use text and image representations of posts which can help describe different aspects of the posts which are not considered in the current model. They are also trying to apply the graph convolutional networks on the brand-mentioned network to improve the performance of the model. They also plan to create a model that can detect paid advertisements which do not contain any sponsorship information.
9.	Andre Maureen Pudjajana, Danny Manongga, Ade Iriani, Hindriyanto Dwi Purnomo 2018 International Seminar on	Identification of Influencers in Social Media using Social Network Analysis (SNA)	This research aims to identify influencers in social media with focus on the measurement of Degree, Closeness, and Betweenness Centrality.	This research uses hoax dataset. Each centrality measurement result is combined with weighting to each of the centrality values in each vertex. Then the result of the	The result of this research is hierarchical model of 2.5% accounts. The established hierarchy model generates a spread path model and support accounts that involved	Weighting and modelling hierarchy results help to identify influencers and find active

	Research of Information Technology and Intelligent Systems (ISRITI). doi:10.1109/isriti.2018.8864458		The results of this study are a list of SNA measurement results which are then combined using weighting.	weighting is modelled into hierarchy	actively in spreading the hoax. This can be a way to eradicate negative influencers in social media.	accounts in the graph. These models can be looked upon by organizations .
10.	Shubhi Goel , Dr. R.K. Dwivedi and Anu Sharma. SMART–2020, IEEE Conference ID: 50582 9th International Conference on System Modeling & Advancement in Research Trends, 4th–5th, December, 2020.	Analysis of Social Network using Data Mining Techniques	The authors proposed to build a model to understand how 'opinions' about a certain topic get formed. Twitter tweets has been chosen. In their model of the world, an opinion has two elements: Abstraction: What the opinion is about, and Expression: The 'sentiment' of the opinion, i.e. positive, negative or neutral.	The methodology followed is: 1. Getting tweets discussing a specific theme. 2. Cleaning the tweets (eliminating stop words and soon.) 3. Computing embeddings of tweets using pretrained word embedding's (such aswor2vec). 4. Computing tweet embeddings with sentiment by appending a 'sentiment score' to the 'tweet embedding's. 5. Clustering the tweet-sentiment embedding's . 6. Identifying key characteristic phrases of each narrative using TF-IDF frequencies of commonly occurring phrases. 7. Finally, compute the sentiment score	This paper turns into the foundation for future dynamic and anticipating any associations.	The model generated can be further used by organizations to know more about users and help in their business models.

3. DATA DESCRIPTION-

We have taken a dataset from Kaggle platform. It is basically a graph dataset containing various data on over 500 Facebook pages of organizations which are directly or indirectly related to football.

Link-

https://www.kaggle.com/tanmaymisra/football-facebook-pages-dataset

4. METHODOLOGY-

Using the data from the dataset, we will be generating different statistics and graph statistical measures using the network of these pages as a graph structure.

These statistics include -

1) aggregating pages based on their category.
2) top pages based on their fan count (likes).
3) top pages based on total people talking about them.
4) top pages based on page posting activity.
5) correlation between fans & talking about for pages.
6) Graph properties such as-
   a) diameter
   b) edge density
   c) transitivity
   d) coreness
   e) centrality measures i.e. Betweenness, Closeness and Eigenvector
7) PageRank
8) finding communities or clusters in the network
9) modularity score.

We will be producing visualizations which will greatly aid the user in understanding the statistics and this can provide organizations or businesses to better know their fanbase which can help in targeted marketing.

4.1 Framework, Architecture or Module of the Proposed System

The architecture of the proposed system consists of code chunks, each containing some code which can run independently of other chunks. I have outlined the modules and given their brief description below –

1. Importing the required libraries – gridExtra, igraph and ggplot2.
2. Reading in the graph and viewing its brief summary
3. Inspecting the page graph object
4. Aggregating pages based on their category
5. Top pages based on their fan count (likes)
6. Top pages based on total people talking about them
7. Top pages based on page posting activity
8. Checking correlation between fans and talking_about for pages:

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. An intelligent correlation analysis can lead to a greater understanding of your data.

9. Plotting page network using degree filter
10. Getting filtered graph

The filtered_graph class template is an adaptor that creates a filtered view of a graph. The predicate function objects determine which edges and vertices of the original graph will show up in the filtered graph. If the edge predicate returns true for an edge then it shows up in the filtered graph, and if the predicate returns false then the edge does not appear in the filtered graph. Likewise for vertices.

11. Plotting the graph
12. Diameter (length of longest path) of the network
13. Getting the longest path of the network
14. Mean distance between 2 nodes in the network

15. Distance between various important pages (nodes)

16. no. of edges / no. of all possible edges

17. transitivity clustering coefficient

In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterized by a relatively high density of ties; this likelihood tends to be greater than the average probability of a tie randomly established between two nodes.

18. The k-core of a graph is a maximal subgraph in which each vertex has at least degree k.

19. The coreness of a vertex is k if it belongs to the k-core but not to the (k+1)-core.

20. Max coreness

21. Viewing the core of the network

22. Viewing the periphery of the network

(a) *We study a model of network formation where the benefits from connections exhibit decreasing returns and decay with network distance. We show that the unique equilibrium network is a periphery-sponsored star, where one player, the center, maintains no links and earns a high payoff, while all other players maintain a single link to the center and earn lower payoffs.*

(b) *Both the star architecture and payoff inequality are preserved in an extension of the model where agents can make transfers and bargain over the formation of links, under the condition that the surplus of connections increases in the size of agents' neighbourhoods. Our model thus generates two common features of social and economic networks: (1) a core-periphery structure; (2) positive correlation between network centrality and payoffs.*

23. Degree centrality

Degree centrality assigns an importance score based purely on the number of links held by each node.

24. Closeness centrality

In a connected graph, the normalized closeness centrality (or closeness) of a node is the average length of the shortest path between the node and all other nodes in the graph. Thus the more central a node is, the closer it is to all other nodes.

25. Betweenness centrality

One of those hidden centralities is the betweenness centrality, where we look at how often a node lies on the shortest path between two other nodes. The more a node appears on one of those shortest paths, the higher its betweenness centrality. A node with high betweenness is also called a broker as it fulfils a brokerage position in the network, which means that information needs to pass through that entity to be shared by the other nodes.

26. Viewing top pages based on above measures

27. Correlation plots

a) Graph 1 - degree vs. closeness

b) Graph 2- degree vs. betweenness

28. Eigenvector Centrality

Like degree centrality, EigenCentrality measures a node's influence based on the number of links it has to other nodes within the network. EigenCentrality then goes a step further by also taking into account how well connected a node is, and how many links their connections have, and so on through the network.

29. PageRank

PageRank is a variant of Eigen Centrality, also assigning nodes a score based on their connections, and their connections' connections. The difference is that PageRank also takes link direction and weight into account – so links can only pass influence in one direction and pass different amounts of influence.

30. Example of finding neighbours of page vertices

31. Get communities / clusters

32. filtering graph to get important nodes based on degree

33. Fast greedy clustering

(a) The graph package implements a variety of network clustering methods, most of which are based on Newman-Girvan modularity.

(b) The simplest such algorithm is the "fast greedy" method, which starts with nodes in separate clusters, and then merges clusters together in a greedy fashion.

34. get total pages in each cluster

35. get page names in each cluster

36. get modularity score

(a) Modularity is one measure of the structure of networks or graphs. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

(b) Modularity is often used in optimization methods for detecting community structure in networks. However, it has been shown that modularity suffers a resolution limit and, therefore, it is unable to detect small communities.

37. edge betweenness clustering

5. PROPOSED SYSTEM MODEL (Mathematical Modeling)-

A mathematical model is an abstract model that uses mathematical language to describe the behaviour of a system.

There are several modules in my system which are direct programming implementations of mathematical concepts, listed as follows –

A) Aggregating pages based on their category –

this module uses sorting the pages based on their category, and displaying the top 10 pages

B) Top pages based on their fan count (likes), total people talking about them, and page posting activity -

this module sorts the pages based on the respective attribute and displays the top 10 pages

C) Correlation between fans and talking_about for pages-

this chunk uses the concept of correlation, which is basically how closely 2 things are related. The R-squared measure for correlation b/w fans and talking_about attributes is 0.659 which denotes that they are positively correlated.

D) Diameter (length of longest path) of the network -

this module simply finds the diameter (longest path) of the network

E) Mean distance between 2 nodes in the network -

this module finds the mean distance b/w 2 nodes in the network

F) Edge density of the graph -

this module uses the edge_density function in igraph, which finds the no. of edges divided by no. of all possible edges in the graph

G) Transitivity clustering coefficient -

this chunk finds the clustering coefficient of the graph

H) k-core of the graph

I) Degree centrality, Closeness centrality, Betweenness centrality

J) Correlation plots – degree vs. closeness & degree vs. betweenness

K) Pagerank

L) Finding communities / clusters in the graph

M) Fast greedy clustering

N) Getting page names in each cluster

O) Getting the modularity score

P) Edge betweenness clustering

5.1 Proposed System analysis-

As mentioned earlier, the main objective of this system is to compute and find different attributes and features related to a graph, by which one can get a deeper understanding of how a social network function in general.

This system is designed in the form of independently executable code chunks, the visualizations produced as a result are also mostly displayed inline (along with the code).

A description of what each chunk contains exactly can be found in the section Framework, Architecture or Module of the Proposed System.

The system contains all major attributes and features of a social network graph including computations of PageRank, Modularity score, Edge betweenness clustering, Fast greedy clustering, correlation plots, degree / betweenness / closeness centralities, etc.

Thus, this system will serve as a good source for someone who wishes to gain a deep understanding of these concepts via a practical approach and provide useful insights that we can generate from such data, can be used by corporates and organizations to improve their business models, and do targeted marketing campaigns.

6. IMPLEMENTATION

This section will contain pictures of the respective outputs for the different graph features which have been calculated in the proposed system.

I) Reading in the graph and viewing its summary

```
pl_graph <- read.graph(file = "pl.gml.txt", format = "gml")

summary(pl_graph)
```

```
IGRAPH db4f4b4 D--- 582 2810 --
+ attr: id (v/n), label (v/c), graphics (v/c), fan_count (v/c),
| category (v/c), username (v/c), users_can_post (v/c), link (v/c),
| post_activity (v/c), talking_about_count (v/c), Yala (v/c), id (e/n),
| value (e/n)
```

II) Inspecting the page graph object

```
pl_df <- data.frame(id = V(pl_graph)$id,
                    name = V(pl_graph)$label,
                    category = V(pl_graph)$category,
                    fans = as.numeric(V(pl_graph)$fan_count),
                    talking_about = as.numeric(V(pl_graph)$talking_about_count),
                    post_activity = as.numeric(V(pl_graph)$post_activity),
                    stringsAsFactors = FALSE )
View(pl_df)
```

id	name	category	fans	talking_about	post_activity
10	Premier League	Sports League	39301910	634081	0.31
20	TAG Heuer	Jewelry/Watches	2823063	30796	0.14
30	Carling	Food & Beverage Company	200508	12078	0.03
40	Hull Tigers	Sports Team	1000560	40500	0.23
50	Middlesbrough FC	Sports Team	431967	25042	0.29
60	Burnley Football Club	Sports Team	352042	3279	0.19

III) Aggregating pages based on their category

```
grid.table(as.data.frame(sort(table(pl_df$category), decreasing = TRUE)[1:10]),
        rows = NULL,
        cols = c('Category', 'Count'))
```

Category	Count
Athlete	151
Sports Team	77
Community	31
Product/Service	26
Non-Profit Organization	21
Company	20
Local Business	16
Games/Toys	15
Sports League	14
Travel Company	13

IV) Top pages based on their fan count (likes)

```
grid.table(pl_df[order(pl_df$fans, decreasing = TRUE),
            c('name', 'category', 'fans')][1:10, ],
        rows = NULL)
```

name	category	fans
Cristiano Ronaldo	Athlete	118925300
FC Barcelona	Sports Team	95491169
Manchester United	Sports Team	72214897
UEFA Champions League	Sports League	60636892
Neymar Jr.	Athlete	59214746
David Guetta	Musician/Band	54789663
Chelsea Football Club	Sports Team	47253090
Nike Football	Product/Service	42498785
Nike Football	Product/Service	42498683
Nike Football	Product/Service	42498679

V) Top pages based on total people talking about them

```
grid.table(pl_df[order(pl_df$talking_about, decreasing = TRUE),
                 c('name', 'category', 'talking_about')][1:10, ],
           rows = NULL)
```

name	category	talking_about
Cristiano Ronaldo	Athlete	3532172
FC Barcelona	Sports Team	1633078
Manchester United	Sports Team	1618239
Chelsea Football Club	Sports Team	1301568
UEFA Champions League	Sports League	1168555
Neymar Jr.	Athlete	1085169
Etihad Stadium	Stadium	1043577
Sergio Ramos	Athlete	1027324
LaLiga	Sports League	970684
Stamford Bridge	Stadium	862744

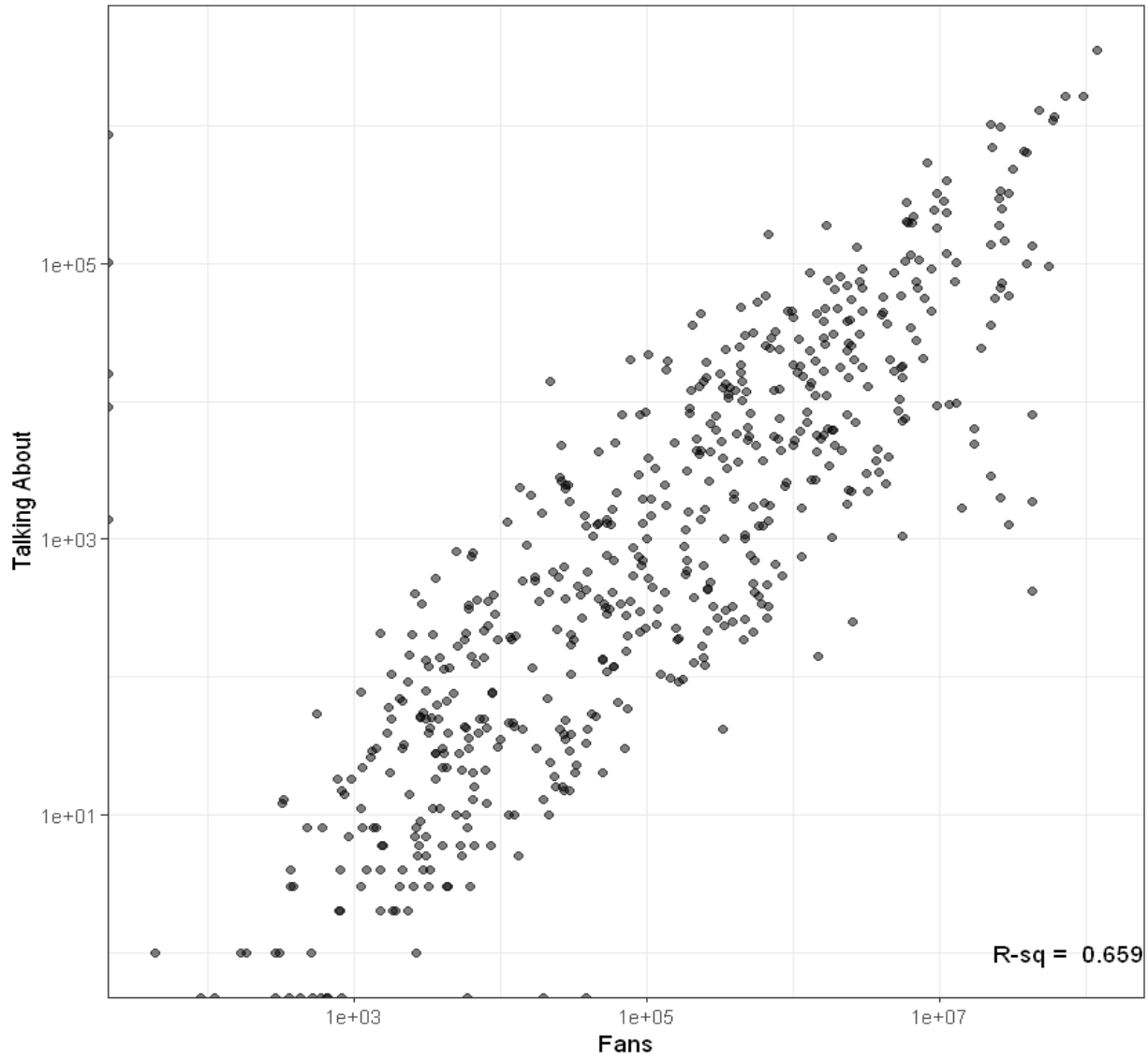
VI) Top pages based on page posting activity

```
grid.table(pl_df[order(pl_df$post_activity, decreasing = TRUE),
                 c('name', 'category', 'post_activity')][1:10, ],
           rows = NULL)
```

name	category	post_activity
SportPesa Care	Product/Service	21.74
GiveMeSport - Football	News/Media Website	10.54
Virgin Media	Telecommunication Company	9.94
Neymar Jr.	Athlete	8.77
Delta	Travel Company	2.83
Juan Mata	Athlete	2.37
The Sims	Games/Toys	2.23
Sky Sports	TV Network	2.18
ESPN UK	Media/News Company	2.02
Sergio Ramos	Athlete	1.67

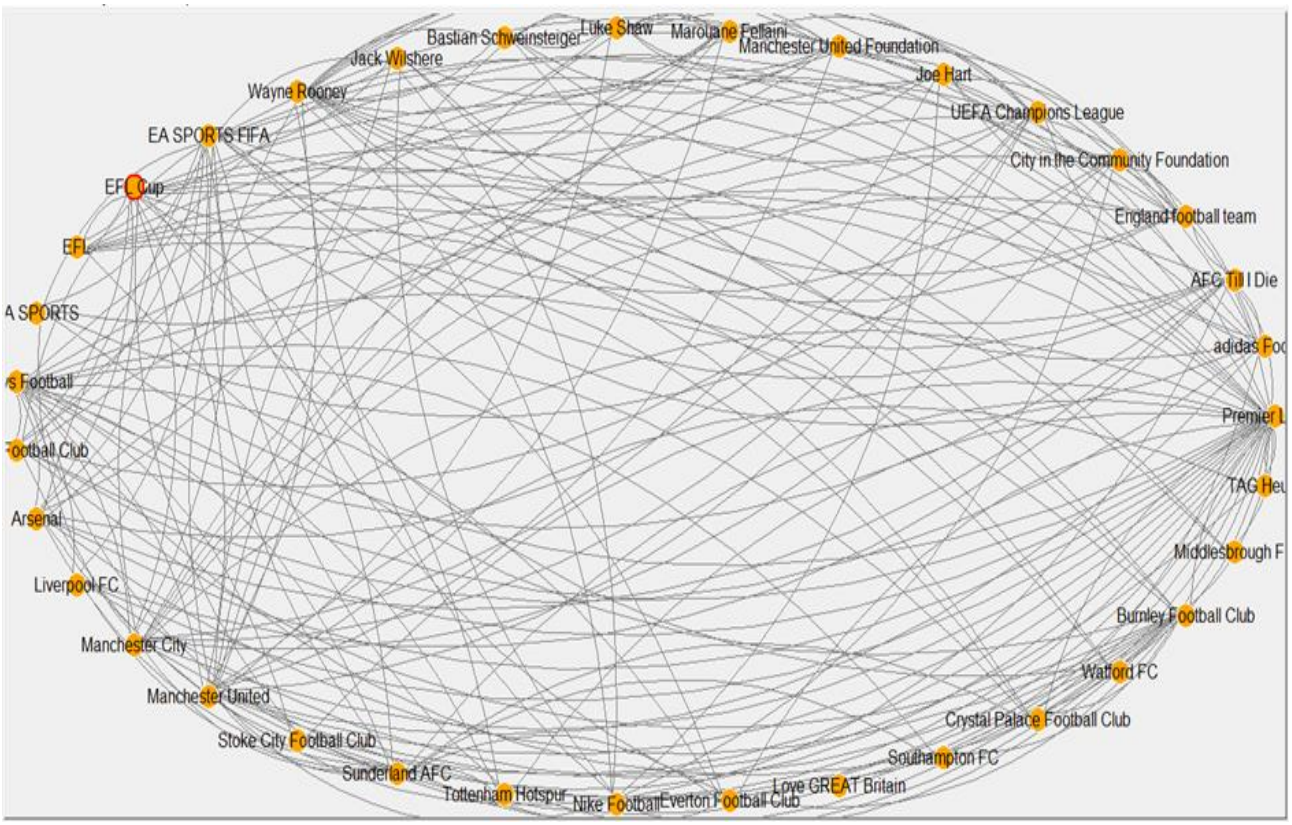
VII) Checking correlation b/w fans and talking_about for pages

```
clean_pl_df <- pl_df[complete.cases(pl_df), ]
rsq <- format(cor(clean_pl_df$fans, clean_pl_df$talking_about) ^ 2, digits = 3)
corr_plot <- ggplot(pl_df, aes(x = fans, y = talking_about)) + theme_bw() +
  geom_jitter(alpha = 1/2) +
  scale_x_log10() +
  scale_y_log10() +
  labs(x = "Fans", y = "Talking About") +
  annotate("text", label = paste("R-sq = ", rsq), x = +Inf, y = 1, hjust = 1)
corr_plot
```



VIII) Plotting page network using degree filter

```
degrees <- degree(pl_graph, mode = "total")
degrees_df <- data.frame(ID = V(pl_graph)$id,
                         Name = V(pl_graph)$label,
                         Degree = as.vector(degrees))
ids_to_remove <- degrees_df[degrees_df$Degree < 30, c('ID')]
ids_to_remove <- ids_to_remove / 10
# Getting filtered graph
filtered_pl_graph <- delete.vertices(pl_graph, ids_to_remove)
# Plotting the graph
tkplot(filtered_pl_graph,
     vertex.size = 10,
     vertex.color = "orange",
     vertex.frame.color = "white",
     vertex.label.color = "black",
     vertex.label.family = "sans",
     edge.width = 0.2,
     edge.arrow.size = 0,
     edge.color = "grey",
     edge.curved = TRUE,
     layout = layout.fruchterman.reingold)
```



IX) Diameter (length of the longest path) of the network

```
diameter(pl_graph, directed = TRUE)
```

7

X) Getting the longest path of the network

```
get_diameter(pl_graph, directed = TRUE)$label
```

'Sports Arena Hull' 'Hull Tigers' 'Teenage Cancer Trust' 'Celtic FC' 'Dafabet UK' 'Premier League' 'Carling' 'Alice Gold'

XI) Mean distance b/w 2 nodes in the network

```
mean_distance(pl_graph, directed = TRUE)
```

3.69602850897396

XII) Distance b/w various important pages (nodes) [an example]

```
node_dists <- distances(pl_graph, weights = NA)
labels <- c("Premier League", pl_df[c(21, 22, 23, 24, 25), 'name'])
filtered_dists <- node_dists[c(1, 21, 22, 23, 24, 25), c(1, 21, 22, 23, 24, 25)]
colnames(filtered_dists) <- labels
rownames(filtered_dists) <- labels
grid.table(filtered_dists)
View(filtered_dists)
```

	Premier League	Manchester United	Manchester City	Liverpool FC	Arsenal	Chelsea Football Club
Premier League	0	1	1	1	1	1
Manchester United	1	0	2	2	2	2
Manchester City	1	2	0	2	2	2
Liverpool FC	1	2	2	0	2	2
Arsenal	1	2	2	2	0	2
Chelsea Football Club	1	2	2	2	2	0

XIII) Edge density of the graph

```
edge_density(pl_graph)
```

```
0.00831011823435125
```

XIV) Transitivity clustering coefficient

```
transitivity(pl_graph)
```

```
0.163949029784267
```

XV) Page coreness (k-core, coreness)

```
page_names <- V(pl_graph)$label
page_coreness <- coreness(pl_graph)
page_coreness_df = data.frame(Page = page_names,
                              PageCoreness = page_coreness)
page_coreness_df


# The k-core of a graph is a maximal subgraph in which each vertex has at least degree k.
# The coreness of a vertex is k if it belongs to the k-core but not to the (k+1)-core.
```

Page	PageCoreness
Premier League	11
TAG Heuer	6
Carling	5
Hull Tigers	9
Middlesbrough FC	7
Burnley Football Club	11
Watford FC	9
AFC Bournemouth	9
Leicester City Football Club	10
Crystal Palace Football Club	11
West Ham United	11
Southampton FC	11
Love GREAT Britain	5
Everton Football Club	11
Nike Football	11
West Bromwich Albion	11
Tottenham Hotspur	11
Swansea City Football Club	11
Sunderland AFC	11
Stoke City Football Club	11
Manchester United	11
Manchester City	11

XVI) Max coreness

```
max(page_coreness_df$PageCoreness)
```

11

XVII) Viewing the core of the network

```
View(head(page_coreness_df[
  page_coreness_df$PageCoreness == max(page_coreness_df$PageCoreness),], 20))
# Viewing the core of the network
```

	Page	PageCoreness
1	Premier League	11
6	Burnley Football Club	11
10	Crystal Palace Football Club	11
11	West Ham United	11
12	Southampton FC	11
14	Everton Football Club	11
15	Nike Football	11
16	West Bromwich Albion	11
17	Tottenham Hotspur	11
18	Swansea City Football Club	11
19	Sunderland AFC	11
20	Stoke City Football Club	11
21	Manchester United	11
22	Manchester City	11
23	Liverpool FC	11
24	Arsenal	11
25	Chelsea Football Club	11
26	Barclays Football	11
77	adidas	11
99	QPR FC	11

XVIII) Viewing the periphery of the network

```
View(head(page_coreness_df[
  page_coreness_df$PageCoreness == min(page_coreness_df$PageCoreness),], 20))
```

	Page	PageCoreness
34	Henrik Lundqvist	1
37	Cara Delevingne	1
38	La Carrera Panamericana	1
39	Patrick Dempsey	1
40	Dempsey Racing	1
52	The Carling Local at V Festival	1
57	Tigers Trust	1
59	Hull Tigers Commercialâ€™	1
60	Hull Tigers Arabic	1
61	Andy Dawson Testimonial	1
84	Safehands Nursery at Barnoldswick	1
88	Liv Fox Photography	1
89	Split Screen Wedding Dreams	1
94	Alex O'Neill Photography	1
95	Pier Fun Casinos Event Management Ltd	1
98	BBC Radio Lancashire	1
104	Myprotein	1
105	Althams Travel Todmorden	1

XIX) Degree centrality

```
degree_plg <- degree(pl_graph, mode = "total")
degree_plg_df <- data.frame(Name = V(pl_graph)$label,
                            Degree = as.vector(degree_plg))
degree_plg_df <- degree_plg_df[order(degree_plg_df$Degree, decreasing = TRUE), ]
View(degree_plg_df)
```

	Name	Degree
22	Manchester City	109
24	Arsenal	92
15	Nike Football	91
19	Sunderland AFC	91
25	Chelsea Football Club	90
1	Premier League	85
21	Manchester United	84
6	Burnley Football Club	82
14	Everton Football Club	65
23	Liverpool FC	59
17	Tottenham Hotspur	58
453	City in the Community Foundation	56
445	UEFA Champions League	53
262	Wayne Rooney	52
26	Barclays Football	51
243	EA SPORTS FIFA	47
416	Marouane Fellaini	47
558	adidas Football	47
12	Southampton FC	45
20	Stoke City Football Club	44

Values in list are in descending order.

XX) Closeness centrality

```
closeness_plg <- closeness(pl_graph, mode = "all", normalized = TRUE)
closeness_plg_df <- data.frame(Name = V(pl_graph)$label,
                               Closeness = as.vector(closeness_plg))
closeness_plg_df <- closeness_plg_df[order(closeness_plg_df$Closeness, decreasing = TRUE), ]
View(closeness_plg_df)
```

	Name	Closeness
1	Premier League	0.5340074
19	Sunderland AFC	0.4874161
6	Burnley Football Club	0.4805624
26	Barclays Football	0.4723577
129	EFL Cup	0.4611111
366	606	0.4469231
243	EA SPORTS FIFA	0.4418251
150	The Offside Rule (We Get It!) Podcast	0.4411541
156	The H and C News Football Pie League	0.4365139
22	Manchester City	0.4342302
21	Manchester United	0.4326136
183	Mondogoal	0.4278351
15	Nike Football	0.4259531
24	Arsenal	0.4234694
242	Kitbag	0.4234694
25	Chelsea Football Club	0.4228530
559	Beyond Sport	0.4182865
14	Everton Football Club	0.4158912
108	Prostate Cancer UK	0.4158912
126	The Emirates FA Cup	0.4114731

Values in list are in descending order.

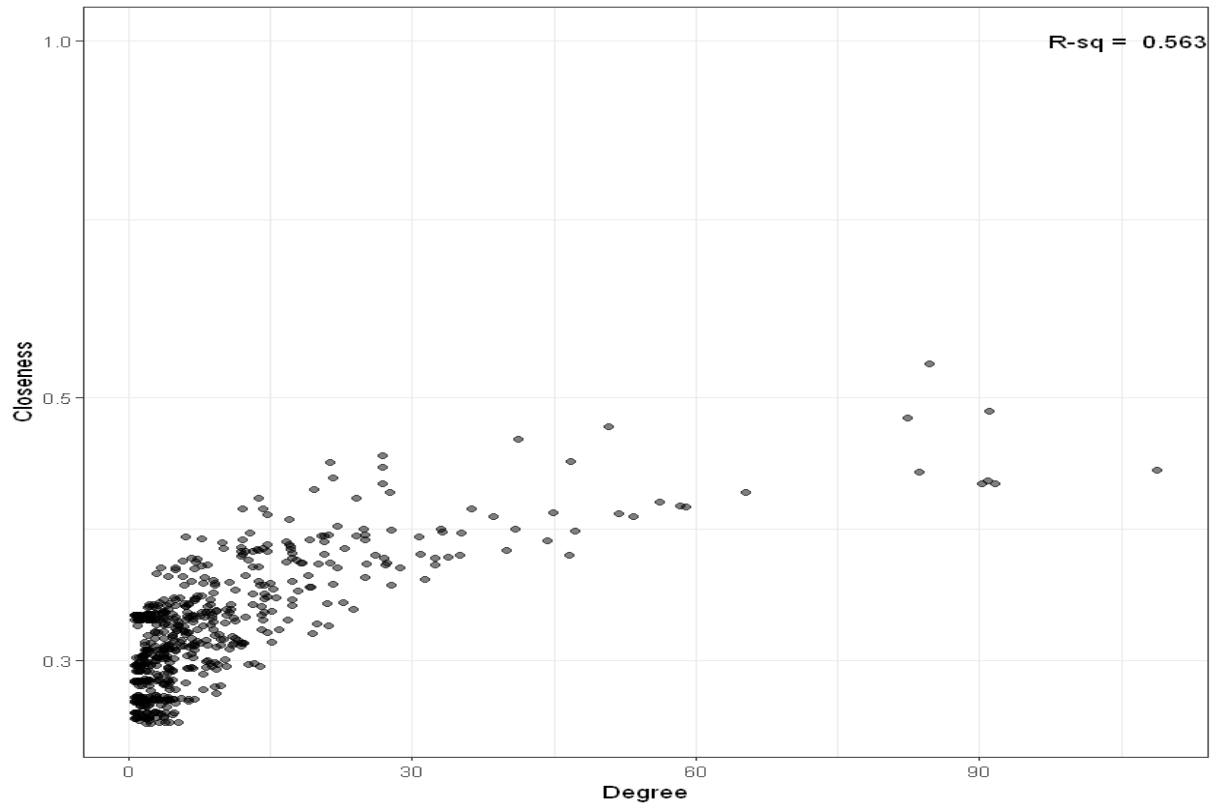
XXI) Betweenness centrality

```
betweenness_plg <- betweenness(pl_graph)

betweenness_plg_df <- data.frame(Name = V(pl_graph)$label,

                                 Betweenness = as.vector(betweenness_plg))

betweenness_plg_df <- betweenness_plg_df[order(betweenness_plg_df$Betweenness, decreasing = TRUE), ]

View(betweenness_plg_df)
```

	Name	Betweenness
1	Premier League	96082.154
22	Manchester City	32242.448
19	Sunderland AFC	24583.192
24	Arsenal	22258.680
6	Burnley Football Club	21872.563
15	Nike Football	21232.926
21	Manchester United	19781.410
243	EA SPORTS FIFA	19013.004
25	Chelsea Football Club	18181.458
2	TAG Heuer	16698.681
23	Liverpool FC	15675.171
14	Everton Football Club	15651.433
26	Barclays Football	15555.314
17	Tottenham Hotspur	14324.345

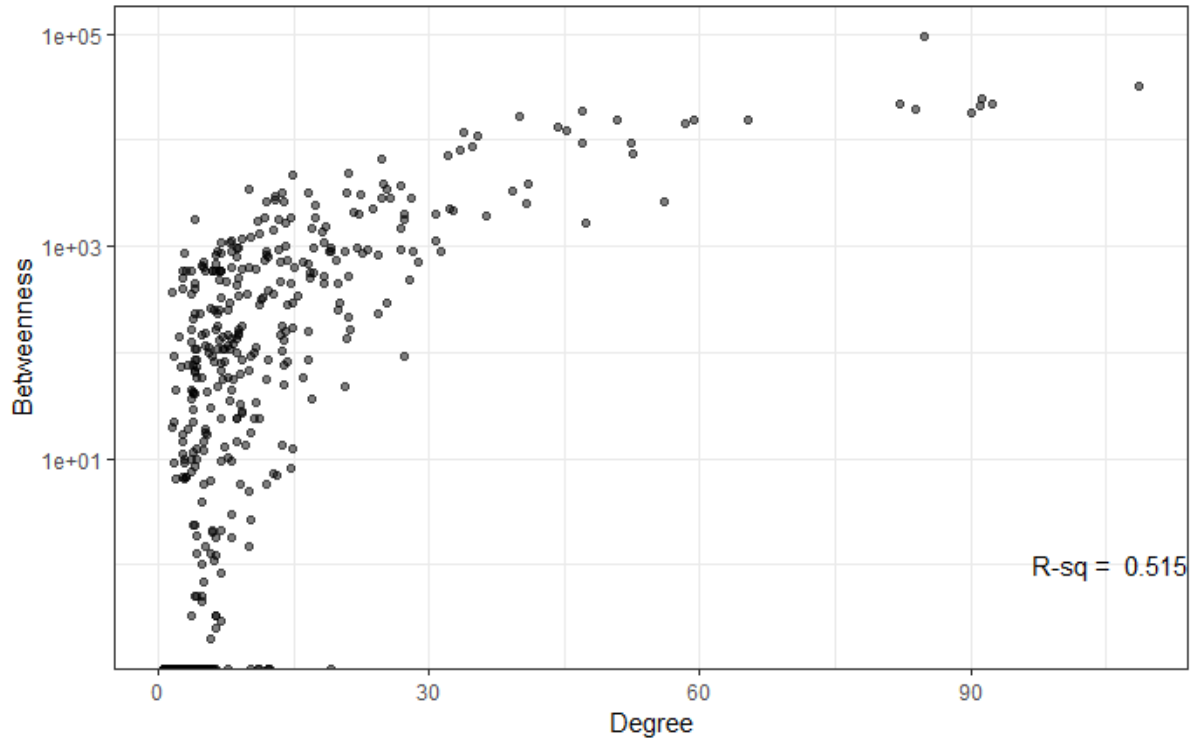
XXII) Correlation plot – degree centrality vs. closeness centrality

```
plg_df <- data.frame(degree_plg, closeness_plg, betweenness_plg)
# Graph 1 - degree vs closeness
rsq <- format(cor(degree_plg, closeness_plg) ^ 2, digits = 3)
corr_plot <- ggplot(plg_df, aes(x = degree_plg, y = closeness_plg)) +
  theme_bw() +
  geom_jitter(alpha = 1/2) +
  scale_y_log10() +
  labs(x = "Degree", y = "Closeness") +
  annotate("text", label = paste("R-sq = ", rsq), x = +Inf, y = 1, hjust = 1)
corr_plot
```



XXIII) Correlation plot – degree centrality vs. betweenness centrality

```
# Graph 2- degree vs betweenness
rsq <- format(cor(degree_plg, betweenness_plg) ^ 2, digits = 3)
corr_plot <- ggplot(plg_df, aes(x = degree_plg, y = betweenness_plg)) +
  theme_bw() +
  geom_jitter(alpha = 1/2) +
  scale_y_log10() +
  labs(x = "Degree", y = "Betweenness") +
  annotate("text", label = paste("R-sq = ", rsq), x = +Inf, y = 1, hjust = 1)
corr_plot
```



PageRank

```
pagerank_plg <- page_rank(pl_graph)$vector
pagerank_plg_df <- data.frame(Name = V(pl_graph)$label,
                              PageRank = as.vector(pagerank_plg))
pagerank_plg_df <- pagerank_plg_df[order(pagerank_plg_df$PageRank, decreasing = TRUE), ]
View(head(pagerank_plg_df, 10))
```

XXIV) Example of finding neighbours of page vertices

```
pl_neighbours <- neighbors(pl_graph, v = which(V(pl_graph)$label == "Southampton FC"))

pl_neighbours
```

```
pl_neighbours$label
```

```
 [1] "Barclays Football"
 [2] "The Emirates FA Cup"
 [3] "JÃ©rÃ©my Pied"
 [4] "Virgin Media"
 [5] "Under Armour (GB, IE)"
 [6] "Radhi JaÃ¯di"
 [7] "Oriol Romeu Vidal"
 [8] "NIX Communications Group"
 [9] "JosÃ© Fonte"
[10] "OctaFX"
[11] "Florin Gardos"
[12] "Harrison Reed"
[13] "Ryan Bertrand"
[14] "Garmin"
[15] "James Ward-Prowse"
[16] "Benali's Big Race"
[17] "Southampton Solent University - Official"
[18] "Sparsholt Football Academy"
[19] "Saints Foundation"
```

XXV) Getting communities or clusters

```
# cliques use different sizes here and experiment

clique_num(pl_graph)
```

```
 [1] "Manchester United"
 [2] "Wayne Rooney"
 [3] "Juan Mata"
 [4] "Bastian Schweinsteiger"
 [5] "David De Gea"
 [6] "Luke Shaw"
 [7] "Daley Blind"
 [8] "Chevrolet FC"
 [9] "Marouane Fellaini"
[10] "Manchester United Foundation"


 [1] "Manchester United"
 [2] "Adnan Januzaj"
 [3] "Wayne Rooney"
 [4] "Juan Mata"
 [5] "Bastian Schweinsteiger"
 [6] "David De Gea"
 [7] "Luke Shaw"
 [8] "Daley Blind"
 [9] "Marouane Fellaini"
[10] "Manchester United Foundation"
```

```
# filtering graph to get important nodes based on degree


degrees <- degree(pl_graph, mode = "total")
degrees_df <- data.frame(ID = V(pl_graph)$id,
                         Name = V(pl_graph)$label,
                         Degree = as.vector(degree_plg))
ids_to_remove <- degrees_df[degrees_df$Degree < 30, c('ID')]
ids_to_remove <- ids_to_remove / 10
filtered_pl_graph <- delete.vertices(pl_graph, ids_to_remove)
fplg_undirected <- as.undirected(filtered_pl_graph)
```

XXVI) Fast greedy clustering

```
fgc <- cluster_fast_greedy(fplg_undirected)
layout <- layout_with_fr(fplg_undirected,
                         niter = 500,
                         start.temp = 5.744)
communities <- data.frame(layout)
names(communities) <- c("x", "y")
communities$cluster <- factor(fgc$membership)
communities$name <- V(fplg_undirected)$label
```

Fast greedy clustering -> total pages in each cluster

```
table (communities$cluster)
```

Output-

```
 1  2  3
15 10 10
```

XXVII) Get page names in each cluster

```
community_groups <- unlist(lapply(groups(fgc),
                          function(item) {
                           pages <- communities$name[item]
                           i <- 1;
                           lim <- 4;
                           s <- ""
                           while(i <= length(pages)) {
                            start = i
                            end = min((i + lim - 1), length(pages))
                            s <- paste(s, paste(pages[start:end], collapse = ", "))
                            s <- paste(s, "\n")
                            i = i + lim
                           }
                           return(substr(s, 1, (nchar(s) - 2)))
                          }))
grid.table(community_groups)
```

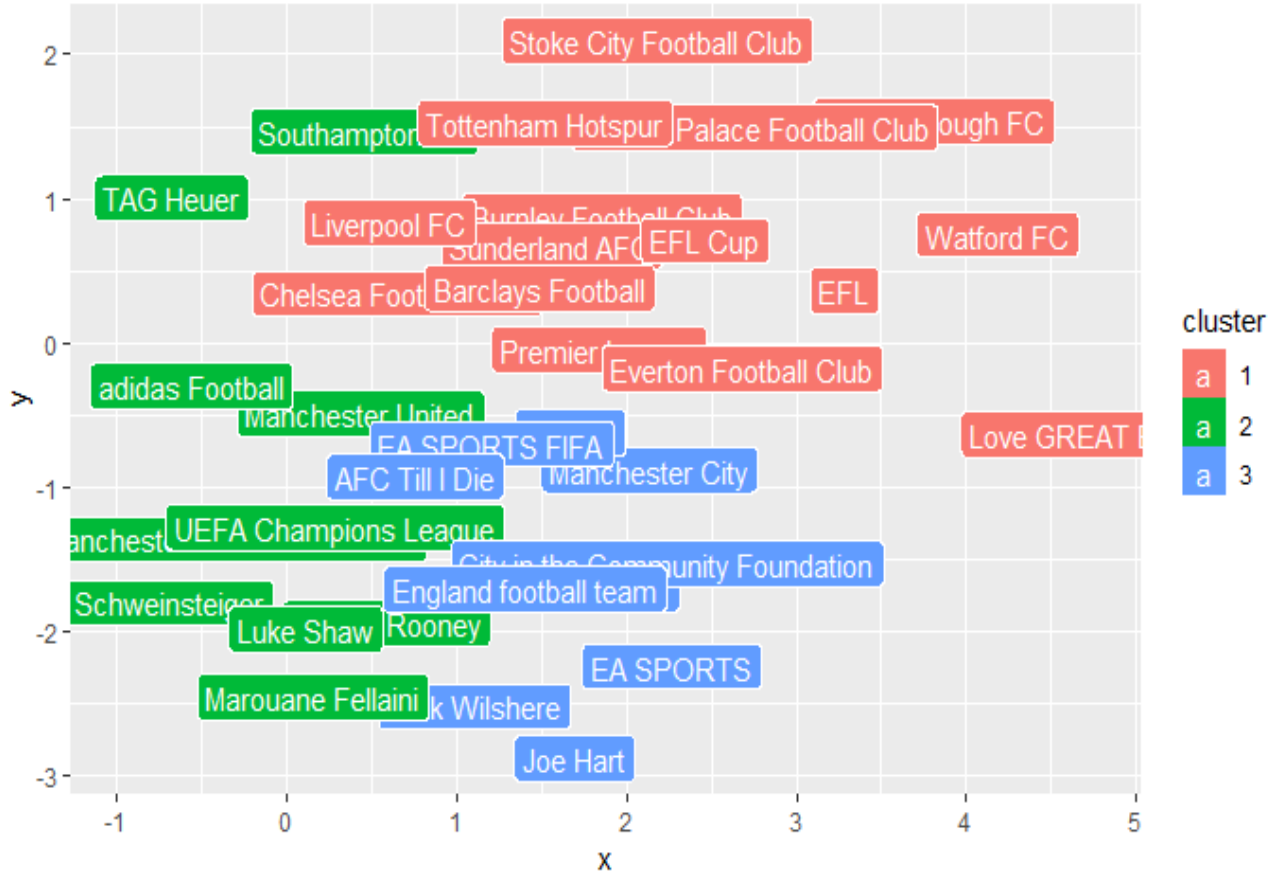
Premier League, Middlesbrough FC, Burnley Football Club, Watford FC Crystal Palace Football Club, Love GREAT Britain, Everton Football Club, Tottenham Hotspur Sunderland AFC, Stoke City Football Club, Liverpool FC, Chelsea Football Club Barclays Football, EFL, EFL Cup
TAG Heuer, Southampton FC, Manchester United, Wayne Rooney Bastian Schweinsteiger, Luke Shaw, Marouane Fellaini, Manchester United Foundation UEFA Champions League, adidas Football
Nike Football, Manchester City, Arsenal, EA SPORTS EA SPORTS FIFA, Jack Wilshere, Joe Hart, City in the Community Foundation England football team, AFC Till I Die

XXVIII) Modularity score

```
modularity(fgc)
```

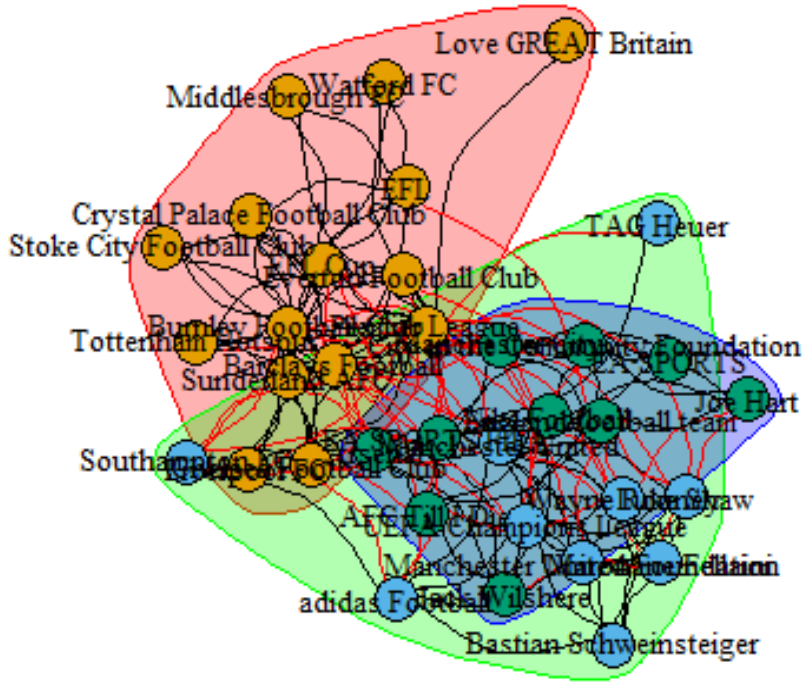
```
[1] 0.2917283
```

```
comm_plot <- ggplot(communities, aes(x = x, y = y, color = cluster, label = name))
comm_plot <- comm_plot + geom_label(aes(fill = cluster), colour = "white")
comm_plot
```



plotting

```
plot(fgc, fplg_undirected,
    vertex.size = 15,
    vertex.label.cex = 0.8,
    vertex.label = fgc$names,
    edge.arrow.size = 0,
    edge.curved = TRUE,
    vertex.label.color = "black",
    layout = layout.fruchterman.reingold)
```

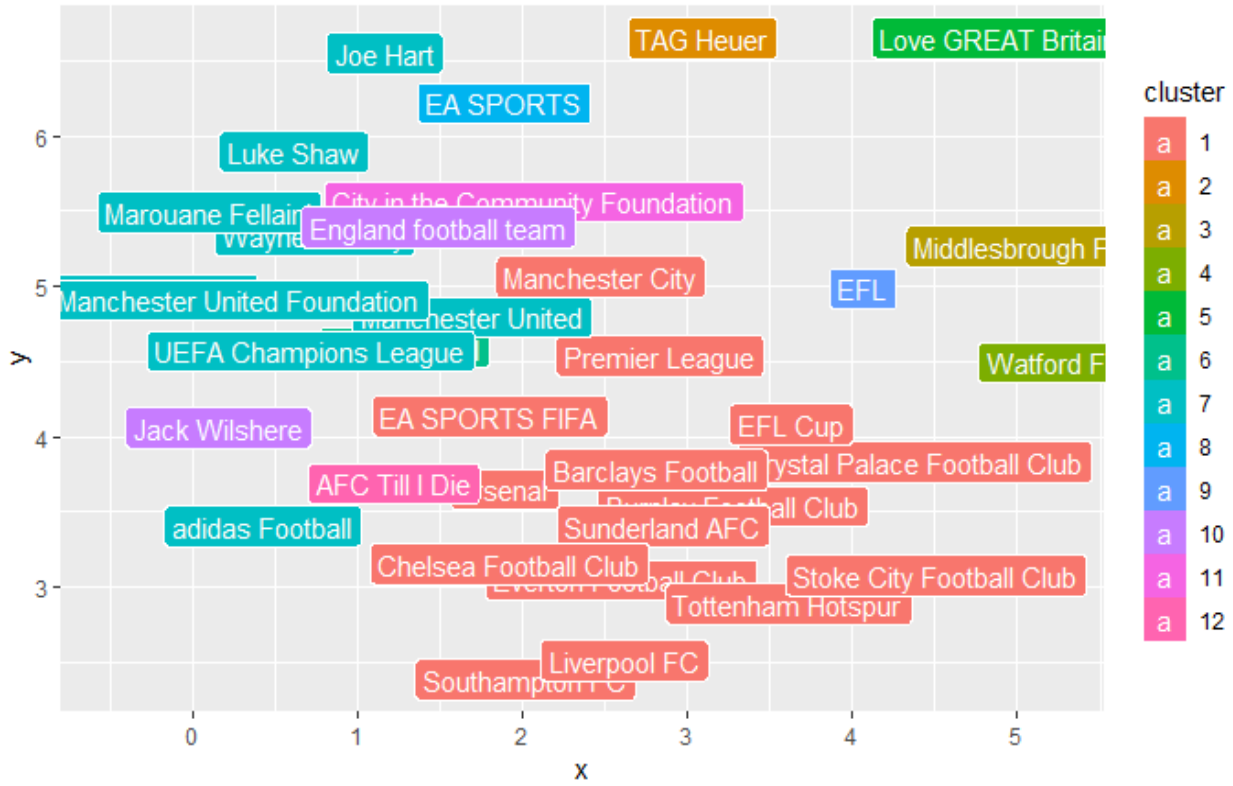


XXIX) Edge betweenness clustering

```
ebc <- cluster_edge_betweenness(fplg_undirected)
layout <- layout_with_fr(fplg_undirected,
                         niter = 500, start.temp = 5.744)
communities <- data.frame(layout)
names(communities) <- c("x", "y")
communities$cluster <- factor(ebc$membership)
communities$name <- V(fplg_undirected)$label
table(communities$cluster)
```

```
 1  2  3  4  5  6  7  8  9 10 11 12
15  1  1  1  1  1  9  1  1  2  1  1
```

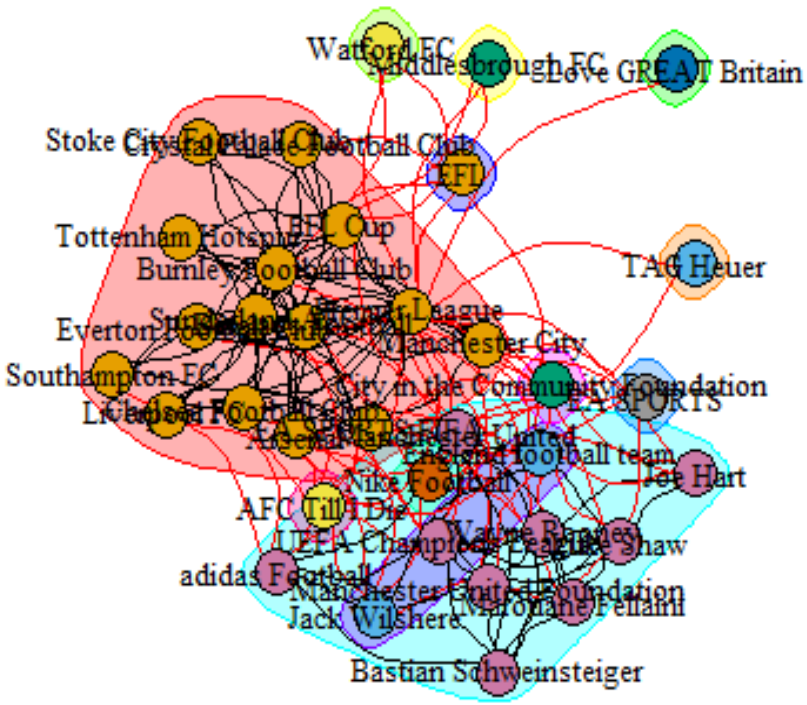
```
comm_plot <- ggplot(communities, aes(x = x, y = y, color = cluster, label = name))
comm_plot <- comm_plot + geom_label(aes(fill = cluster), colour = "white")
comm_plot
```



plotting

```
plot(ebc, fplg_undirected,
    vertex.size = 15,
    vertex.label.cex = 0.8,
    vertex.label = ebc$names,
    edge.arrow.size = 0,
    edge.curved = TRUE,
    vertex.label.color = "black",
    layout = layout.fruchterman.reingold)
```

Output-



7. RESULTS AND DISCUSSION

1) By reading in the graph and viewing its summary, we see that the graph contains 582 rows (pages) and 6 columns – id, name, category, fans, talking_about, post_activity

2) By aggregating pages based on their category, we see that athletes account for the highest no. of pages among all of them, being 151 in number, followed second by ‘Sports Team’ having 77 pages, and so on.

3) When we look at the top pages based on their fan count (likes), we find that Cristiano Ronaldo from the Athlete category has the highest no. of fans (~119 million), followed by FC Barcelona (95.5 million) and Manchester United (72 million) from the Sports Team category.

4) When we look at the top pages based on people talking about them, we find that Cristiano Ronaldo from the Athlete category is on the top with 3.5 million people talking about him, followed by FC Barcelona (1.63 million) and Manchester United (1.62 million) from the Sports Team category.

5) When we look at the top pages based on page posting activity, we see that SportsPesa Care from the Product / Service category is on the top with 21.74 posting activity, followed by GiveMeSport – Football (10.54) from the News / Media website category and Virgin Media (9.94) from the Telecommunication Company category.

6) The correlation plot b/w talking_about and fans shows a positive correlation with the R-squared measure being 0.659 indicating that the 2 features are closely related.

7) The diameter of the network is 7, i.e. the maximum no. of nodes which exist b/w any 2 nodes in the network is 7.

8) The maximum coreness of this network is 11 i.e. the highest value of k for k-core is 11.

9) Pages like ‘Premier League’, ‘Burnley Football Club’ and ‘Crystal Palace Football Club’ have 11 as their coreness and thus form the core of the network. Pages like these are important to the network.

10) Pages like ‘Henrik Lundqvist’, ‘Cara Delevingne’ and ‘Patrick Dempsey’ have the least coreness of 1 and thus form the periphery of the network. Pages like these are not important to the network and they generally exist as a outlier in the network diagram.

11) Manchester City has the highest degree centrality of 109, thus it has the most no. of connections in the network (it’s connected to max. no. of other pages)

12) Premier League has the highest closeness centrality of 0.534, thus it tends to have the least average distance from all other nodes.

13) Premier League also has the highest betweenness centrality of 96082, thus it has the highest tendency to come in the path connecting most other nodes.

14) The correlation between degree centrality and closeness centrality & degree centrality and betweenness centrality is positive, and the R-squared measures of these relations are 0.563 and 0.515 respectively. The latter relation, having a lower R-squared measure proves to be a better correlation i.e. degree centrality and betweenness centrality are better correlated than the former relation.

15) Premier League has the highest Eigenvector centrality (1.0), Nike Football has the highest Pagerank (0.0272) and Manchester United has the highest Authority score (1.0).

All the 3 pages above are thus highly important to the graph, as they seem to have significant connections. Manchester United which has the highest authority score, thus seems to be a page which has in-links from many other pages, making it an important authority figure in the network.

15) The maximal clique calculation process yielded 2 cliques having 10 pages each.

16) The fast greedy clustering process yielded 3 clusters, with the 1st cluster having 15 pages, and the rest of the 2 having 10 pages each.

17) The edge betweenness clustering process yielded 12 communities.

8. NOVELTY

Through our project, we can analyse various networks by Social Network Analysis (SNA), which allow us to better understand how individuals are connected, and most importantly, how information flows – this is critical for improving communication and mobilizing knowledge.

Pages like ‘Premier League’, ‘Burnley Football Club’ and ‘Crystal Palace Football Club’ have highest coreness and thus form the core of the network. This information can be used by Facebook to run advertisements on these pages and the advertisement company gets new customers.

Manchester City has the highest degree centrality of 109, thus it has the most no. of connections in the network. It can be used to convey social messages during emergencies.

Clusters and individuals were also grouped by distance from the core groups. It analysed your network of friends based on your friends' relationships to their other friends.

The edge betweenness clustering process yielded 12 communities. These communities can be analysed further by companies to understand people behaviour and help to encourage football and other sport accessories among them.

Through our project of social network analysis, we can measure, map, and visualize relationships and information flows between individuals or in a group.

9. CONCLUSION

During this project, we have learnt lot of new things about social networks and social network analysis in general. We have added to my skillset, the skills of exploring and cleaning a dataset for analysis, choosing which visualizations would be apt when and where, how to select from a vast variety of graph features and attributes and when to apply them, and how to analyze the outputs and outcomes from the computations performed.

We have worked with real-life data, but we know that social network data is highly variable and doesn't stay static, even for a second.

We have covered important topics related to social network analysis, such as clustering, community analysis, centrality measures, graph features such as diameter, mean distance, coreness, etc. to the best of our knowledge.

10. REFERENCES

1. Graph-based Community Detection in Social Networks. Kaing Sabai Phyu, Myat Myat Min. 2019 IEEE,18th International Conference on Computer and Information Science (ICIS).

2. Centrality Analysis of Online Social Network Big Data. By Hui Li.
2018 IEEE 3rd International Conference on Big Data Analysis.

3. Segment-wise Users' Response Prediction based on Activity Traces in Online Social Networks. By - Oksana Severiukhina, Klavdiya Bochenina National Center for Cognitive Research ITMO University, Saint Petersburg, Russia.

4. https://fmsasg.com/socialnetworkanalysis/facebook/

5. https://towardsdatascience.com/how-to-get-started-with-social-network-analysis-6d527685d374

6. https://firstdraftnews.org/articles/how-to-use-network-analysis-to-explore-social-media-and-disinformation/