CS579 – Online Social Network Analysis

Instagram Hashtags Network Analysis

A20435210 - Karthik Gunasekaran

A20456405 - Shweta Metkar

1. Abstract

Project implements the creation and visualization of a Complex Network of hashtags (referred as tags) on Instagram posts about the IPL cricket for RCB and CSK teams. The project works with some concepts about the process of data collection and construction, handling and visualization of Graphs.

2. Implementation

2.1.Data Collection

On Instagram, a powerful feature to promote ideas in posts are the hashtags, short words prepended by a "#" symbol which express core ideas and feelings about the contents. We easily found the posts which contains tags IPL, RCB, CSK with a simple URL *https://www.instagram.com/explore/tags/* . We have appended */?__a=1* to get the page information in JSON format. Instagram does not give us millions of posts at once. Instead, we need to recursively go down along the page to load more posts. So, a new page is appended to the current one, and we can visualize new posts. This loading process is controlled by a parameter called *end_cursor*, updated at each time a new load is requested.

We designed the Snowballing function to get the posts data. This function returns a list of dictionaries, each one containing the main post information. The depth parameter means how many pages we will dive into. Sometimes, Instagram blocked requests temporarily to prevent the server overload. We used the *sleep* parameter to control the request frequency, *pause* parameter to wait if block fails and *forever* tells the function to wait the pause time and endlessly resume the operation. We got the following numbers of posts for each of these tags:

```
Querying IPLfor depth of 4 pages...
0 1 2 3
Total Number of posts 276
Finished querying for IPL

Now waiting for 30 seconds before querying for the next tag.
Querying CSKfor depth of 4 pages...
0 1 2 3
Total Number of posts 284
Finished querying for CSK

Now waiting for 30 seconds before querying for the next tag.
Querying RCBfor depth of 4 pages...
0 1 2 3
Total Number of posts 270
Finished querying for RCB

Now waiting for 30 seconds before querying for the next tag.
CPU times: user 424 ms, sys: 0 ns, total: 424 ms
Wall time: 1min 59s
```

3. Data Visualization

For visualizing graph, list of edges are enough. We received false or invalid hashtags, to filter them we have created function *validate_tag*. We used network and pandas libraries to create, process and export the graph objects. In summary, we have the following structure to the networks:

- The hashtags are the nodes
- The edges are occurrences of two hashtags in a same post
- The weight of an edge is the frequency in which the corresponding pair of hashtags occur in the sampled posts
- The weight of a node is the frequency in which that hashtag occurs in the sample posts

We created two graph objects: one containing edges between all the hashtags and other containing just edges from the initial target hashtags to all others.

3.1.Gephi:

We use Gephi for visualizing the graphs. The generated .graphml files after data collections and pre-processing is then used to generate the below visualizations using Gephi.

We generated 3 different graphml files:

A. With all the edges and nodes.

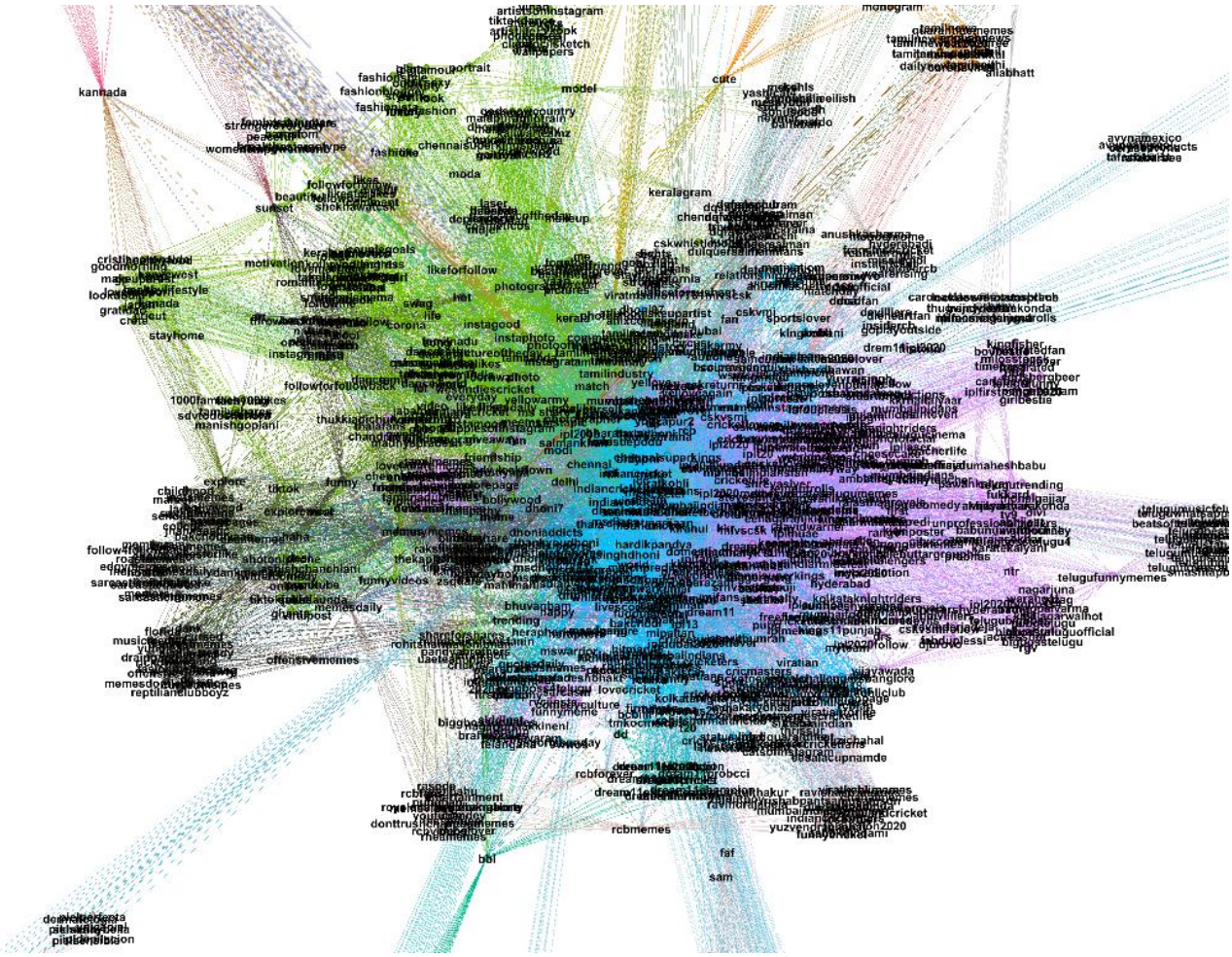
B. With weighted edges.

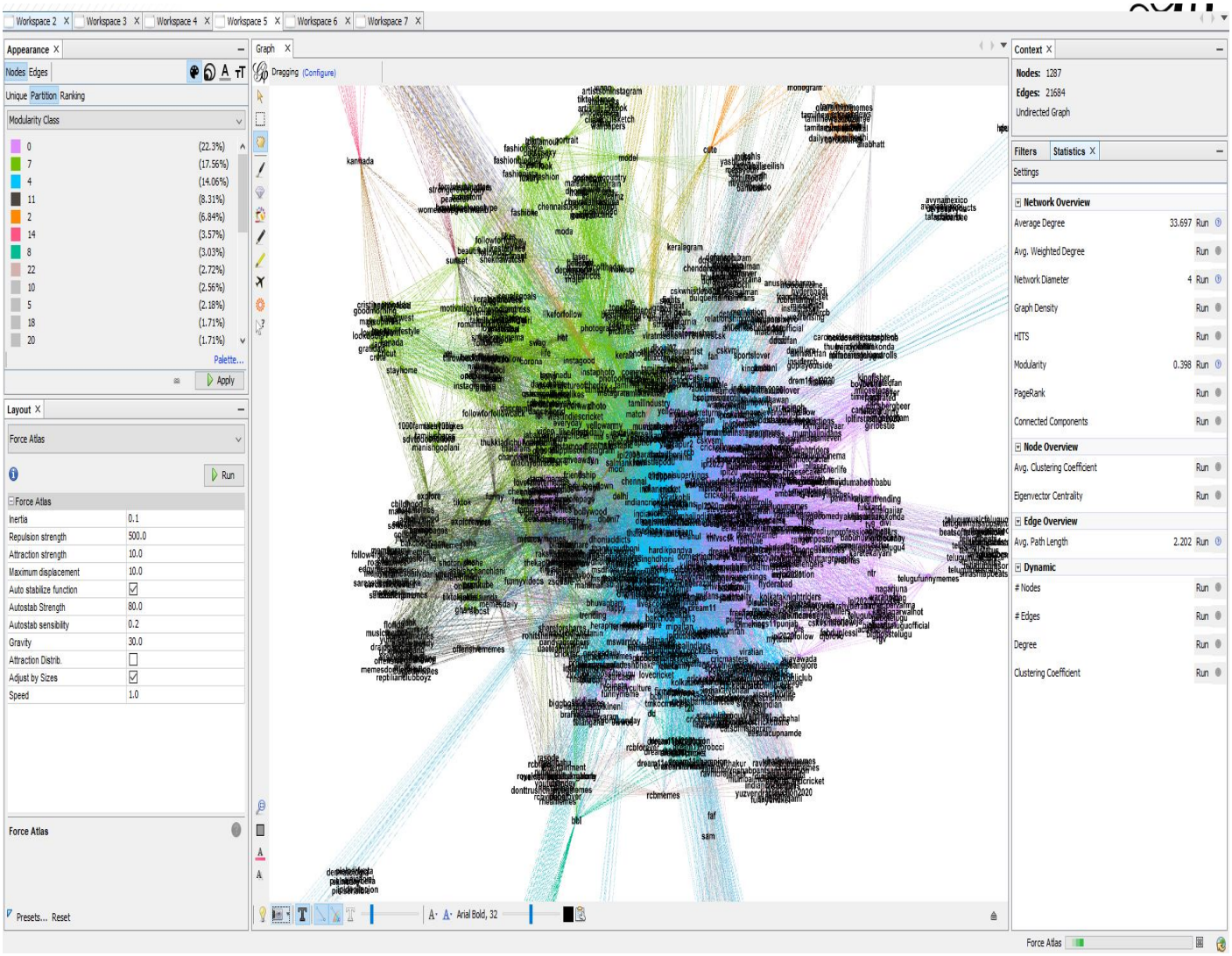
C. With tag nodes (being the centrality).

We created the visualizations using the Force Atlas layout with degree – Between centrality and partitioning the nodes based on Modularity and then the nodes were grouped by sizes.

3.1.1. With all edges and nodes - This graph plots all the edges and nodes from all the posts. It also shows the relation and interconnectivity between each post with multiple hashtags.

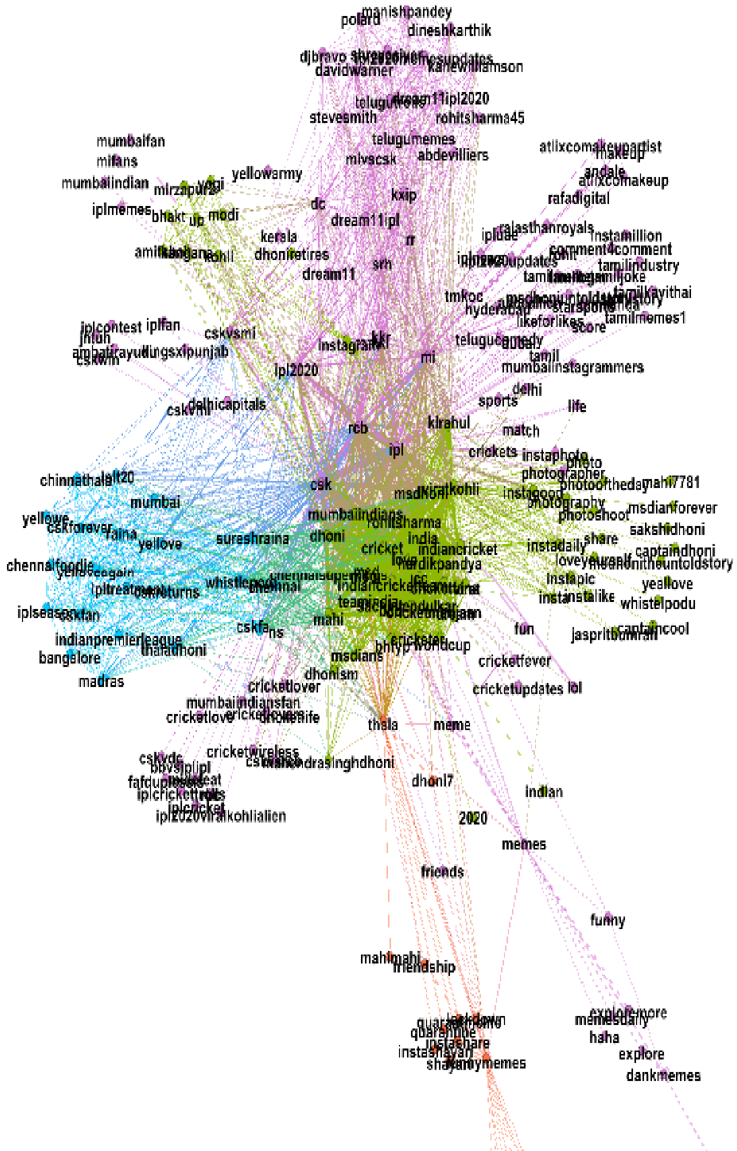
- Number of Nodes – 1287
- Number of Edges – 21684
- Avg Path length – 2.202
- Modularity – 0.398



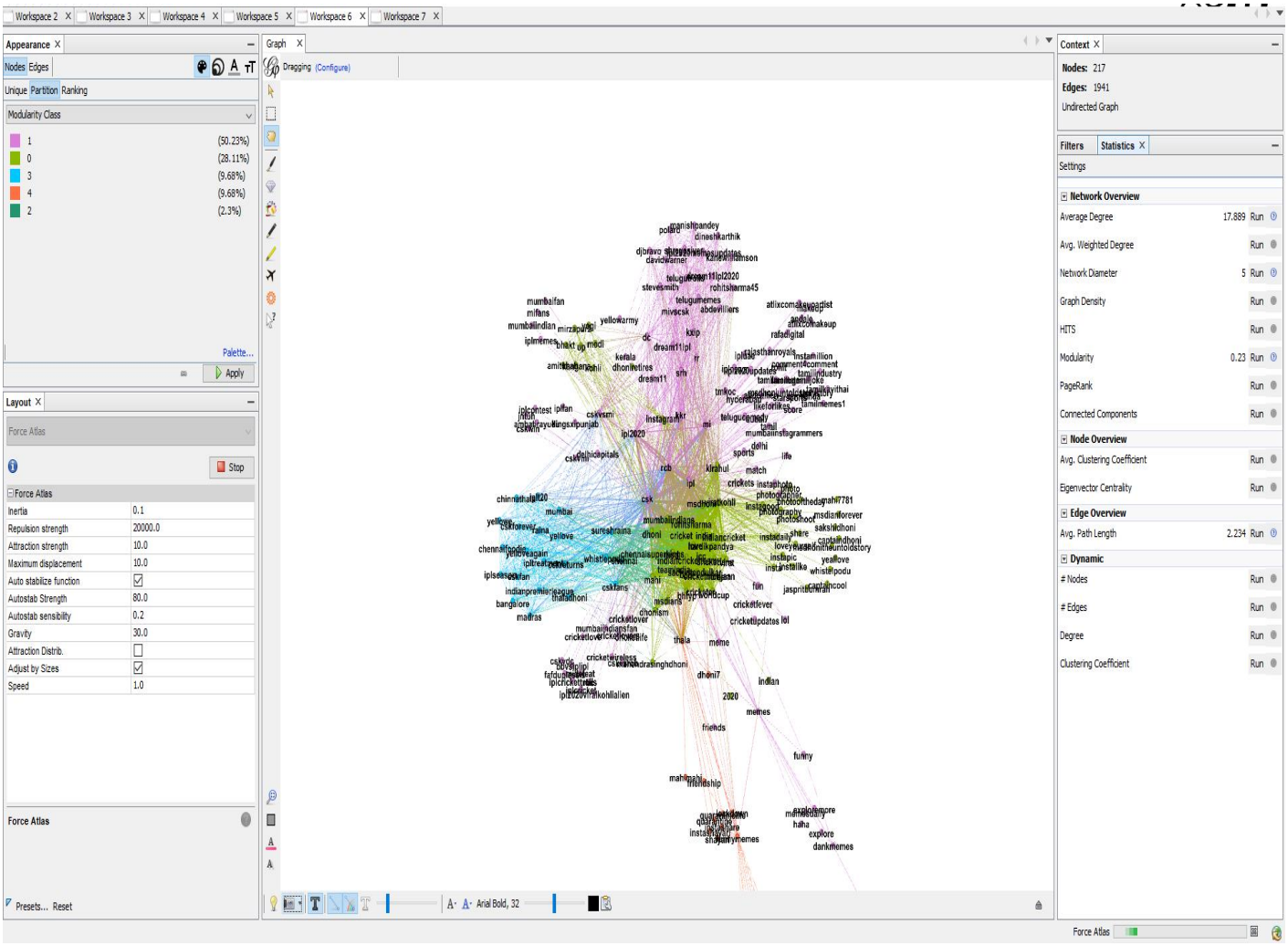


3.1.2. With Weighted edges – This graph only contains edge with weights above the threshold value. The lower weighted edges are ignored and include only the important edges.

- Number of Nodes – 217
- Number of Edges – 1941
- Avg Path length – 2.234
- Modularity – 0.23



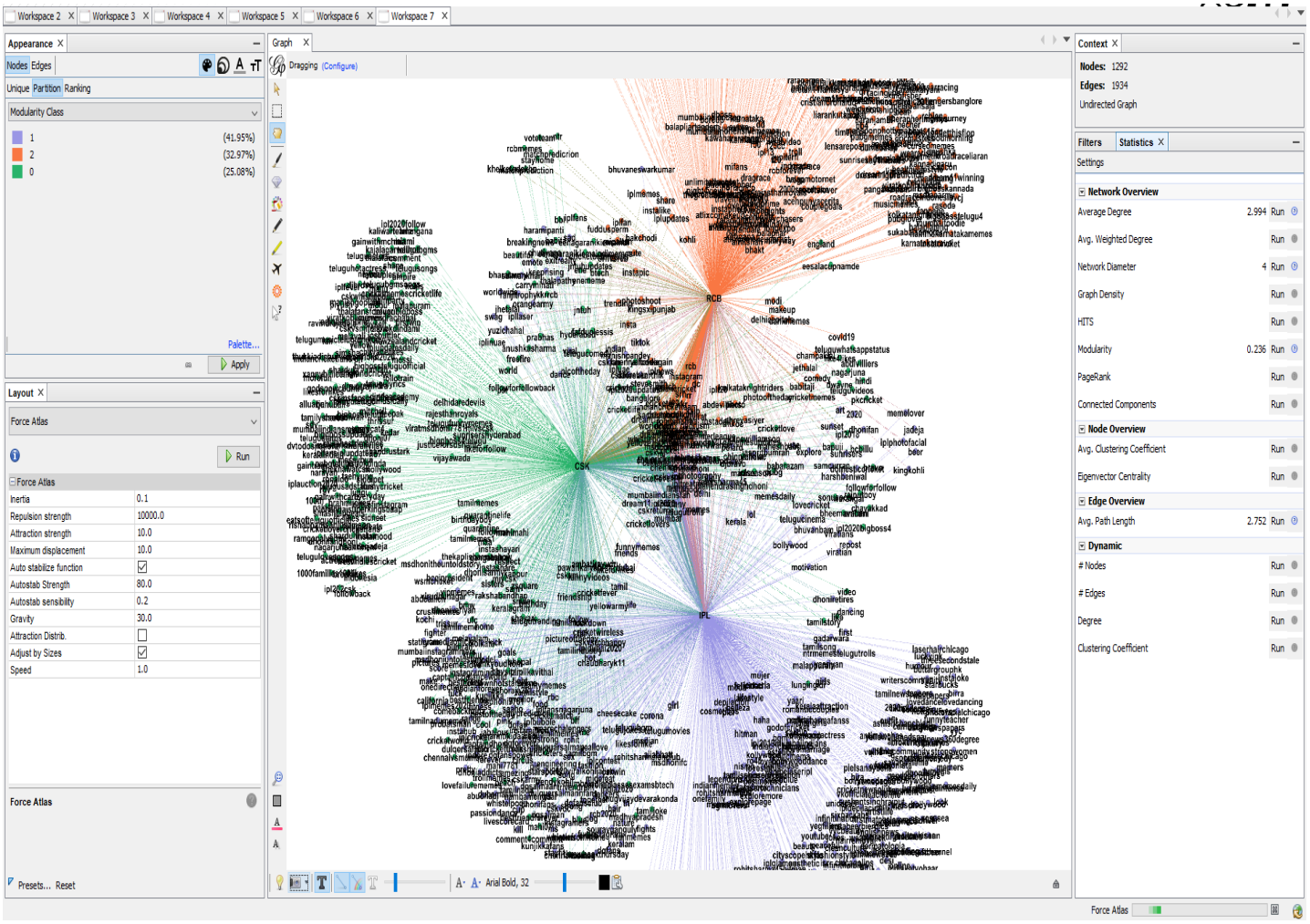
manishpandey
pollard
dineshkarthik
djbravo
kanewilliamson
davidwarner
dream11ipl2020
rohitsharma45
stevesmith
telugumemes
mivscsk
abdevilliers
kxip
dream11ipl
rr
srh
dream11
mumbaifan
mifans
mumbaiindian
iplmemes
yellowarmy
yogi
bhakt
up
modi
dc
kerala
dhoniretires
atlixcomakeupartist
makeup
atlixcomakeup
rafadigital
rajasthanroyals
instamillion
comment4comment
tamilindustry
tamilkavithai
tamilmemes1
tmkoc
hyderabad
likeforlikes
score
teluguwedy
tamil
mumbaiinstagrammers
delhi
sports
life
match
iplcontest
iplfan
cskvsmi
ipl2020
delhicapitals
instagram
kkr
mi
rcb
klrahul
crickets
ipl
instaphoto
photographer
photo
photooftheday
mahi7781
msdhoni
viratkohli
instagood
photography
photoshoot
msdianforever
sakshidhoni
chinnathala
csk
mumbai
mumbaiindians
rohitsharma
yellowe
cskforever
raina
yellove
sureshraina
dhoni
cricket
india
indiancricket
instadaily
share
captaindhoni
hardikpandya
loveyourself
msdhonitheuntoldstory
instapic
yeallove
chennaifoodie
whistlepodu
chennai
teamindia
insta
instalike
whistelpodu
iplseason
indianpremierleague
thaladhoni
cskfans
mahi
fun
jaspritbumrah
captaincool
bangalore
madras
msdians
dhonism
cricketers
worldcup
cricketfever
cricketupdates
lol
cricketlover
mumbaiindiansfan
cricketlove
cricketlife
thala
meme
cricketwireless
mahendrasinghdhoni
dhoni7
indian
2020
memes
friends
funny
friendship
explore
haha
dankmemes



3.1.3. With Tag Nodes- This graph shows how all the posts are related to the main tags. Different colors help understand that posts contains only one of the tags else the posts contains multiple tags and how the tags are connected/relationship with each other.

- Number of Nodes – 1292
- Number of Edges – 1934
- Avg Path length – 2.275
- Modularity – 0.236


RCB
CSK
IPL



4. Network Measures Analysis

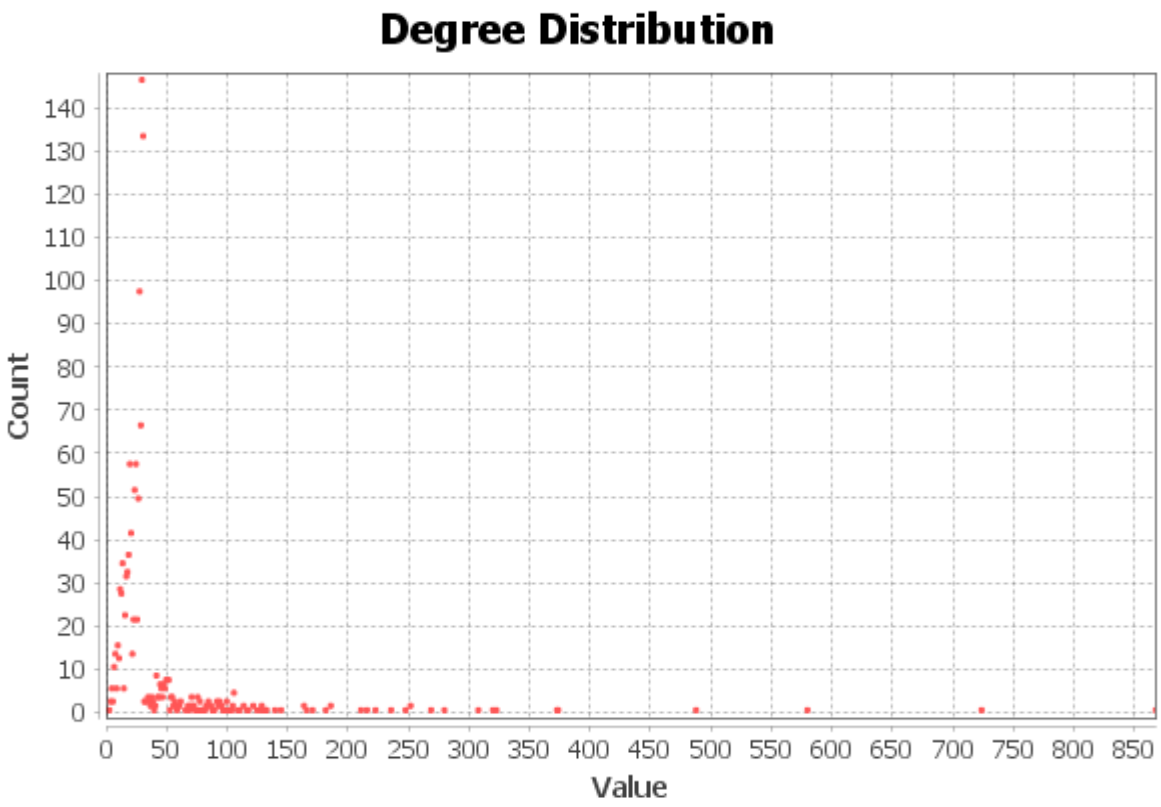
4.1 Degree Distribution:

In the study of graphs and networks, the degree of a node in a network is the number of connections it has to other nodes and the degree distribution is the probability distribution of these degrees over the whole network.

We calculated the degree distribution for the above 3 graphs as follows:

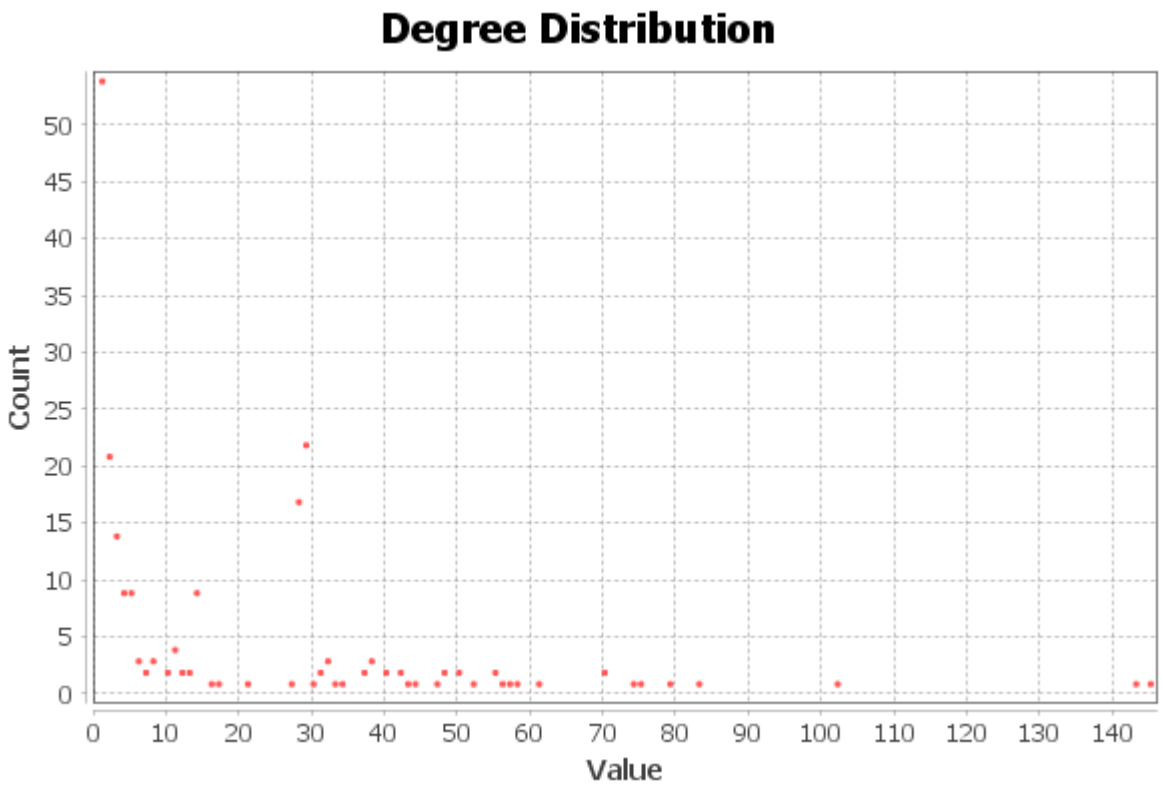
A. With all edges and nodes:

- Avg Degree Distribution: 33.697



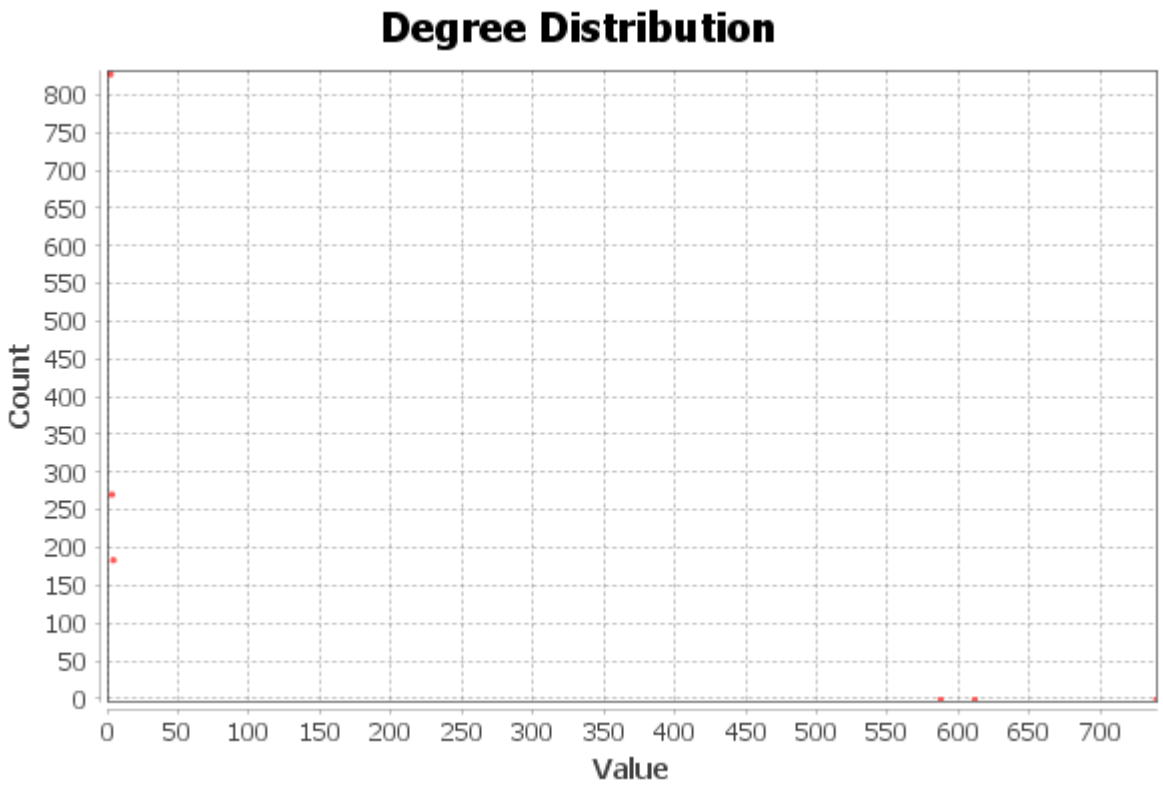
B. With Weighted Edges:

- Avg Degree Distribution: 17.889



C. With Tag Nodes:

- Avg Degree Distribution: 2.994



5. Conclusions

In this project we developed algorithms to collect data from posts on Instagram based on hashtags and constructed and visualized Complex Networks with the data about them. Also, we had limitations of time and hardware to process larger amounts of data, and the graphs generated contains much trash. Maybe, with a larger amount of data, this trash can become insignificant and be dropped from the graph, giving space to other highlights and insights.

6. References

- Martino, Francesco & Spoto, Andrea. (2006). Social Network Analysis: A brief theoretical review and further perspectives in the study of Information Technology. PsychNology Journal. 4. 53-86.
- Manikonda, Lydia & Hu, Yuheng & Kambhampati, Subbarao. (2014). Analyzing User Activities, Demographics, Social Network Structure and User-Generated Content on Instagram.
- Ibba, Simona & Orrù, Matteo & Pani, Filippo & Porru, Simone. (2015). Hashtag of Instagram: From Folksonomy to Complex Network. 279-284. 10.5220/0005613502790284.
- Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.
- Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008
- McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).
- Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, Mart, & Vrgo
  c, Domagoj. (2016). Foundations of JSON schema. In Proceedings of the 25th International Conference on World Wide Web (pp. 263–273).
- Van Rossum, G., & Drake Jr, F. L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.

Appendix

Task breakdown:

Sl.NO	Tasks	Contributed By
1.	Project Proposal	Karthik/ Shweta
2.	Data Collection/Scrapping	Karthik
2.1	Data Cleaning and preprocessing	Karthik
3	Generating graphml files	
3.1	Generating all edges and nodes graphml file	Karthik
3.2	Generating weighted edges graphml file	Shweta
3.3	Generating tag nodes graphml file	Shweta
4.	Gephi Research	Karthik/Shweta
4.1	Visualization all edges and nodes graph	Karthik
4.2	Visualization weighted edges graph	Shweta
4.3	Visualization tag nodes graph	Shweta
5.	Report	Shweta/Karthik