Social Media Mining Project 1

Network Measures

Scraping the data:

I scrapped the data from arxivscraper. I queried the data for "cs" category from August 1, 2017 to August 1, 2018. This data has around 42 thousand records. In the script I filtered for authors in sub-category "math". After filtering for math category the record count reduces to 7,838.

Total No. Of Nodes = 8232

Total No. Of Edges = 22260

Network Creation:

I used networkx library in python to create the graphs. Node list and edge list with weights is fed to the graph. I first created a network with all the nodes and edges. If any two authors have more than one arxiv in common then weight increases accordingly.

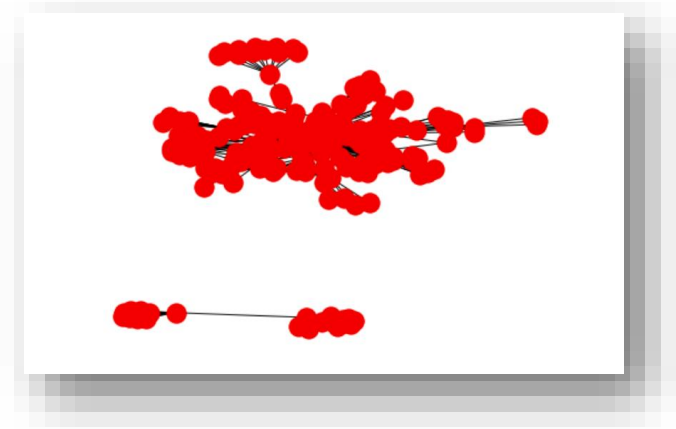
I filtered for nodes with the neighbor count between 10 to 15 and created a list of such nodes(authors). Everytime a node is added to this list all its neighbors are also added. This list is generated as long the count reaches above 200 nodes.

No. Of Nodes after above filtering = 204

No. Of Edges after above filtering = 886

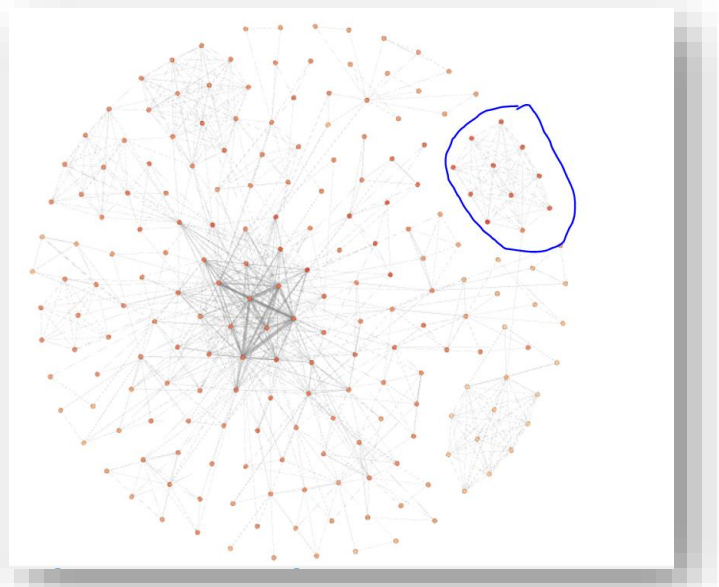
Network Visualization:

Visualizing using matplotlib in python:



Visualizing using Gephi:

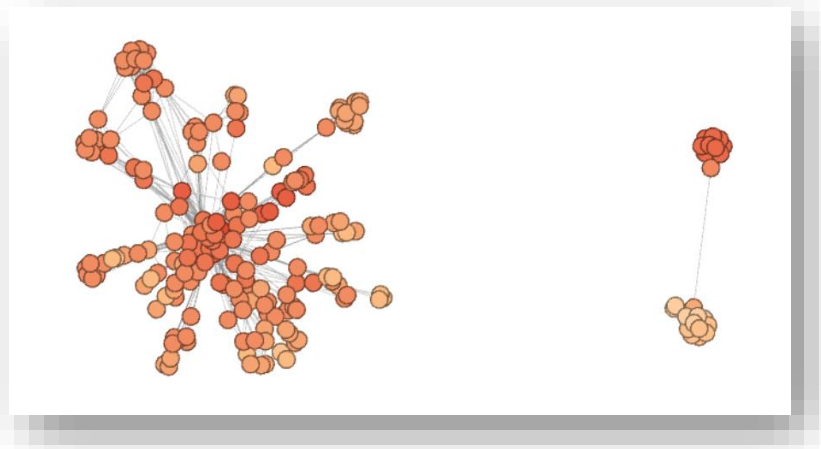
Fruchterman_Reingold view: The highlighted part is the unconnected sub-graph



Force-Atlas View:

Settings: Repulsion Strength = 20.0 Attraction Strength = 2.0

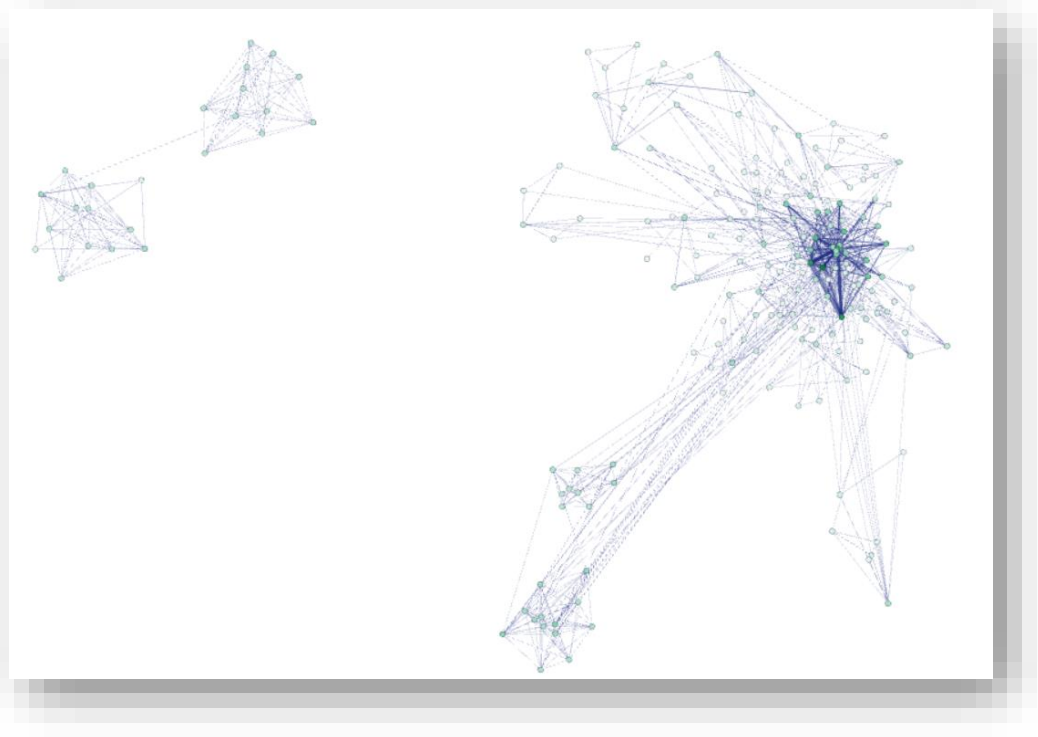
Intensity of node presents the degree.



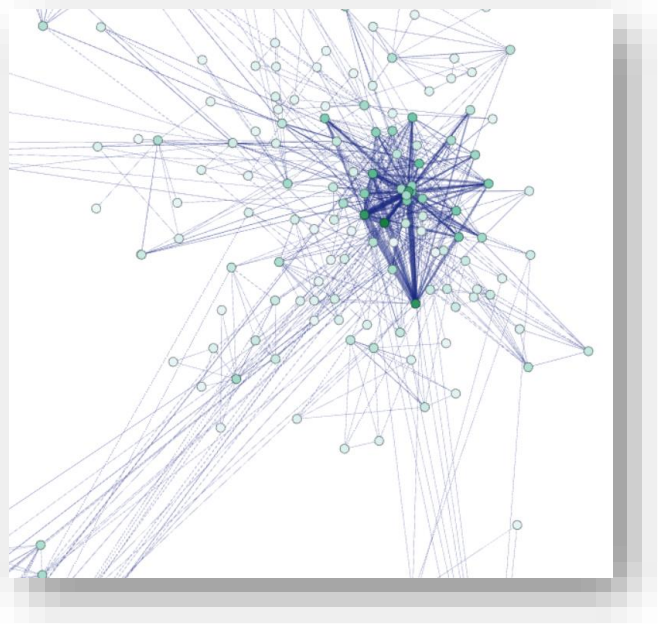
OpenOrd View:

Settings: Default to Gephi

Intensity of node presents the degree. Intensity of edge denotes the weight.



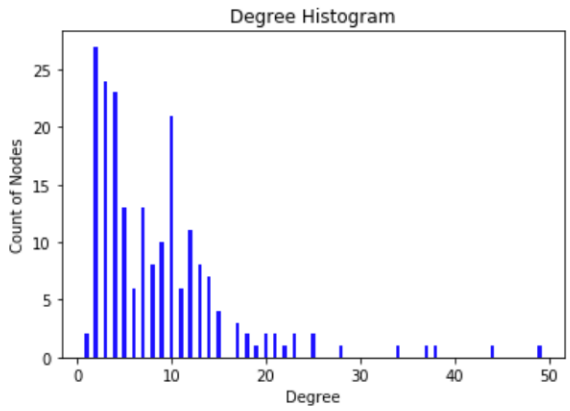
Closer view of dense nodes in sub-graph 1:



Network Measures:

I used networkx package in python to obtain network measures.

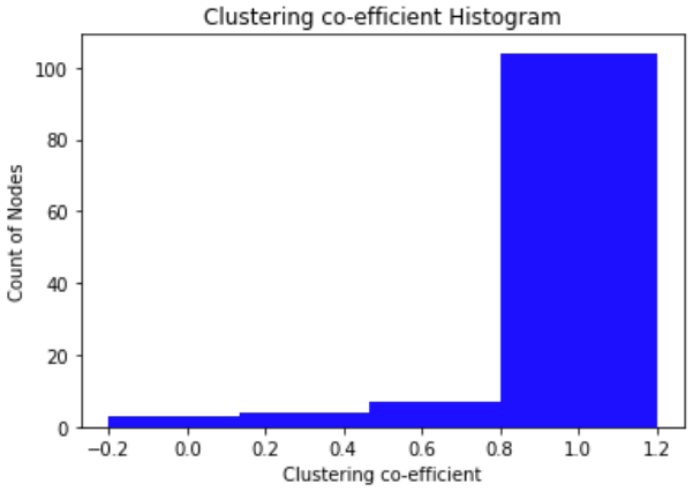
Degree Distribution



Average Degree noted in gephi is 8.686

Clustering Co-efficient

Average Clustering co-efficient of graph is: 0.768721

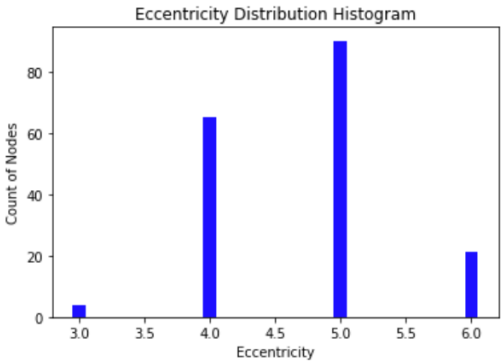


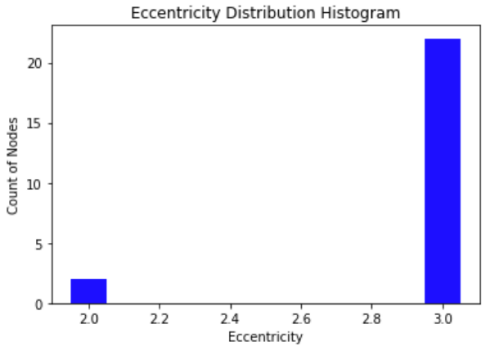
Diameter

Since there are 2 unconnected graphs in the network. Diameter is obtained for each sub-graph:

Diameter of large and small sub-graphs are 6, 3 respectively.

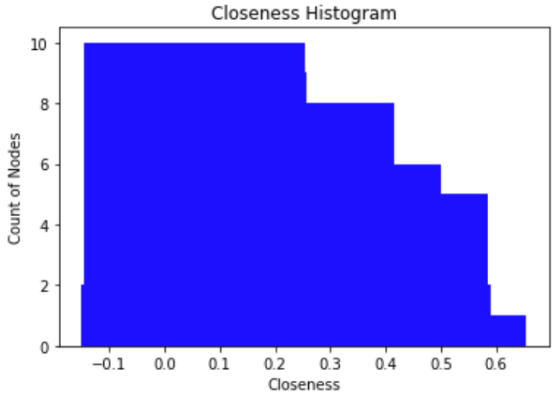
Eccentrcity of each subgraphs are:





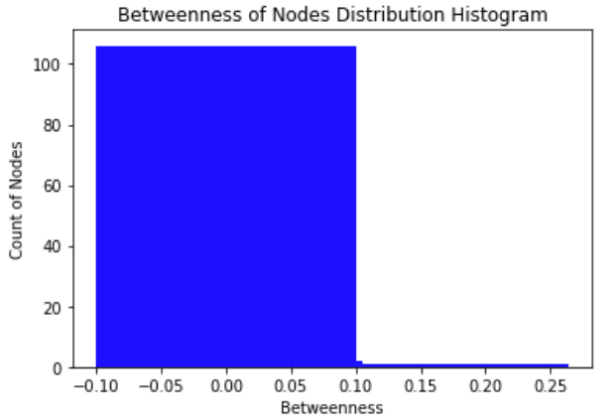
Closeness

Closeness centrality distribution for all nodes is as follows:



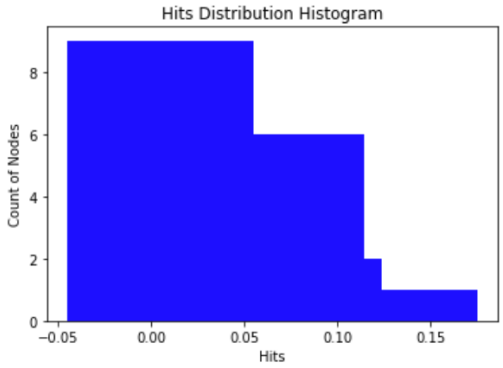
Betweenness

Betweenness centrality distribution for all nodes is as follows:



Hits

Hits distribution for all nodes is as follows:


Hits Distribution Histogram
Count of Nodes
Hits
0
2
4
6
8
−0.05
0.00
0.05
0.10
0.15