[[1]](#footnote-1)

Network Visualization -Week 7/8 Group Project

Sangeetha Verkot, Peter Kim, Rungano Chitakunye

*Abstract*— Visualize various graph datasets using Gephi, generate graph statistics and experiment with layout algorithms.

*Index Terms*—Gephi, Graph, Network, Visualization

# INTRODUCTION

T

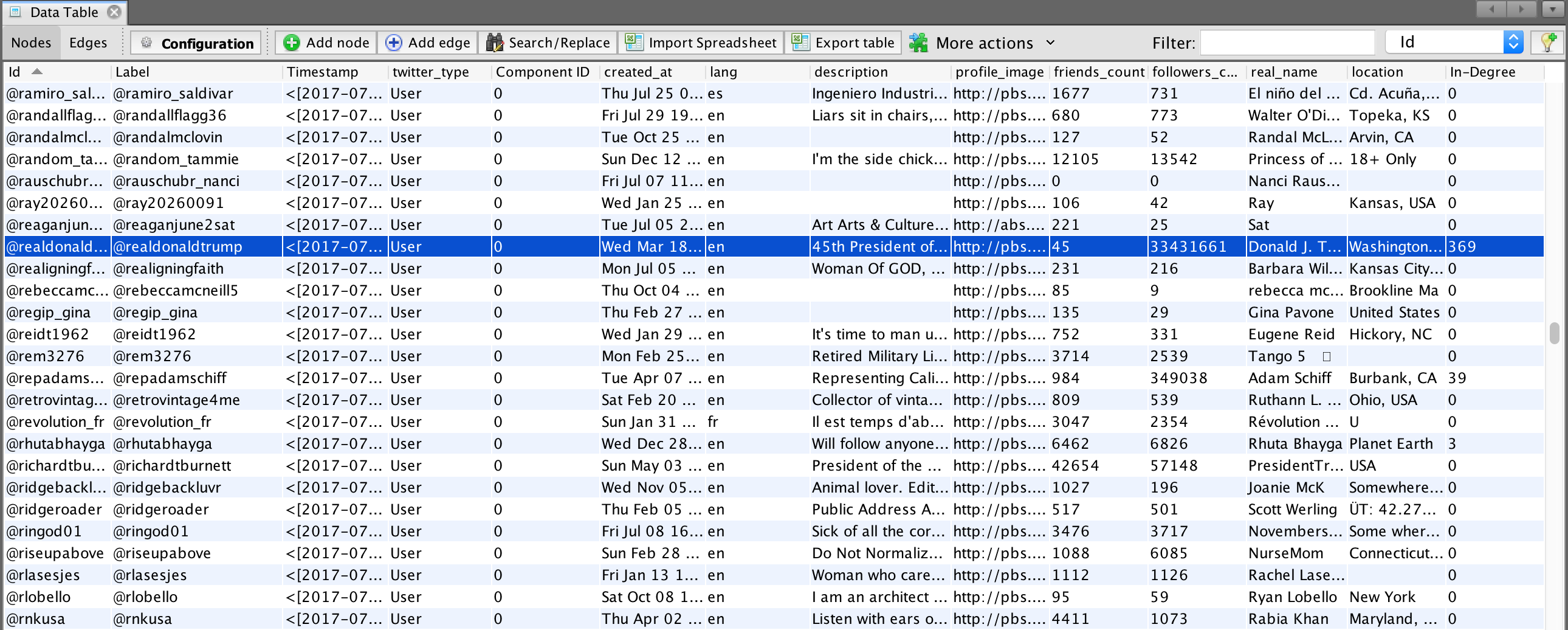
HIS project uses the software Gephi to visualize network data. Gephi has built-in tools to run algorithms like Average Degree, Network Diameter, PageRank, Modularity and generate graph statistic. This data can be analyzed using various layout algorithms such as Fruchterman-Reingold, YifanHu and ForceAtlas to name a few.

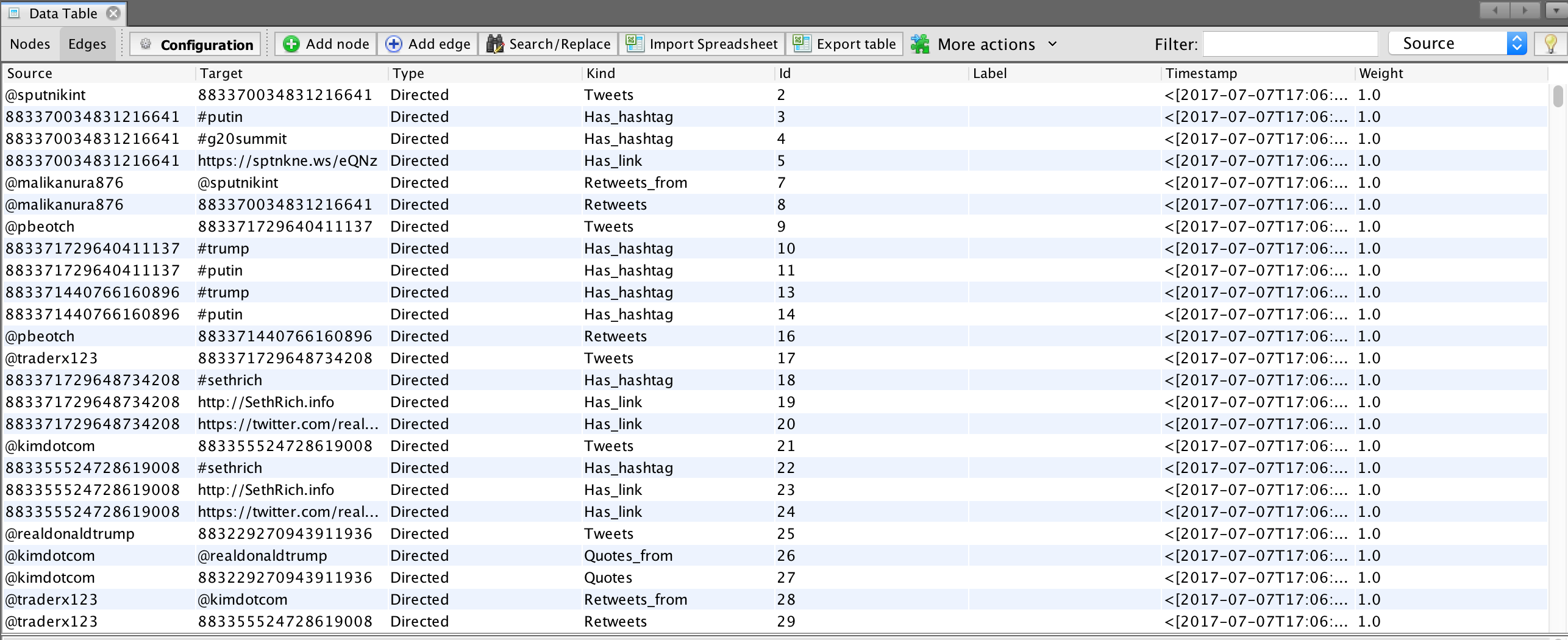
# twitter network analysis

Twitter Streaming Importer is a plug-in that can collect tweets in real-time on the topic of your choice. It will generate a data table and create connections between the users mentioned in those tweets. Since the G-20 summit is happening right now and Donald Trump is a prolific Twitter user, I decided to see how many times @realdonaldtrump was mentioned and within seconds the graph was populated with tweets that mention @realdonaldtrump and I had to disconnect so my system won’t run out of memory.

## Data Table

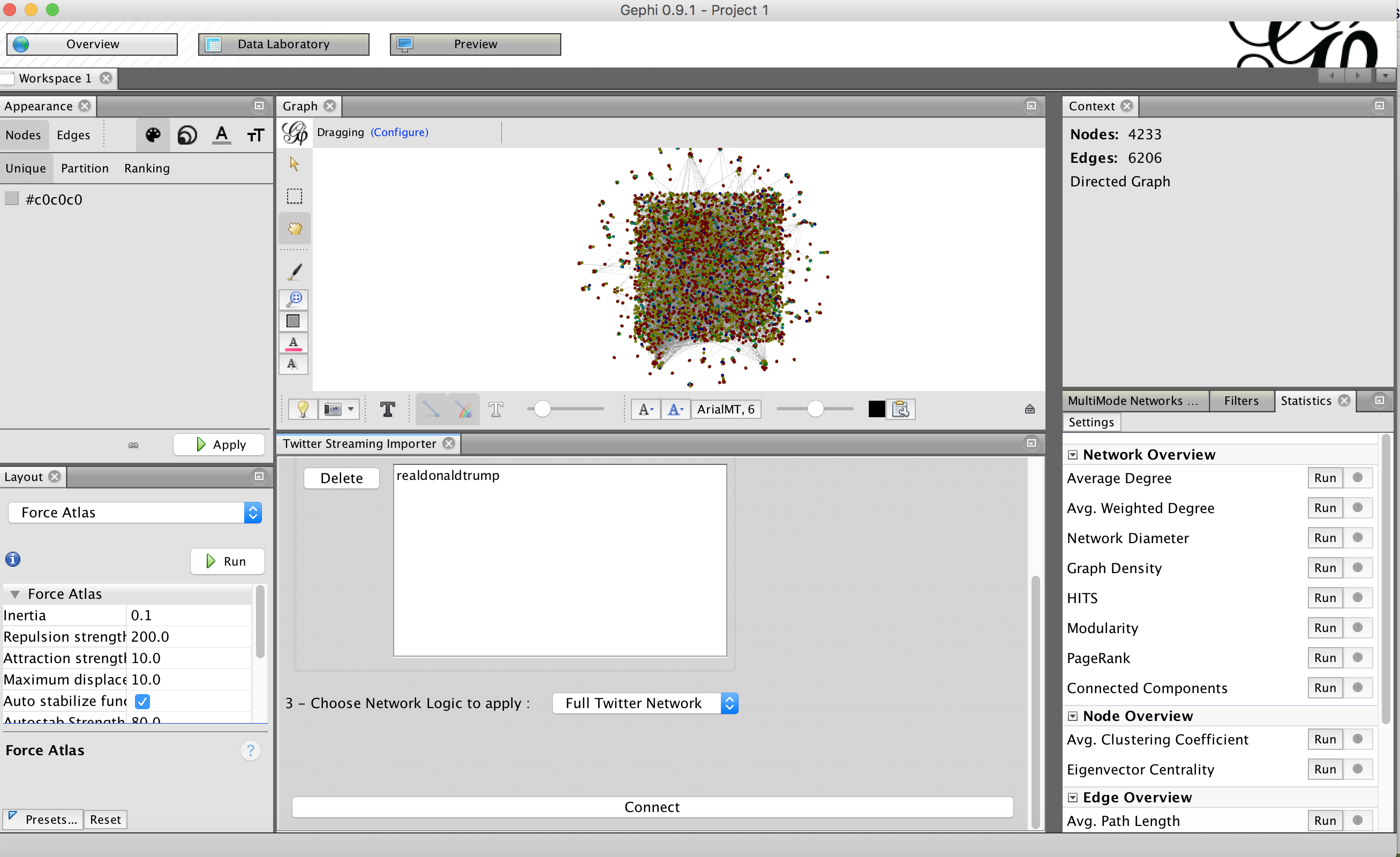
Here’s a snapshot of the Nodes table and the Edges table. It has the twitter handles of the users as the Id the tweets as labels. The connections between the users mentioned in these tweets are used as source and target values in the edges table.





## A First Look at the Graph

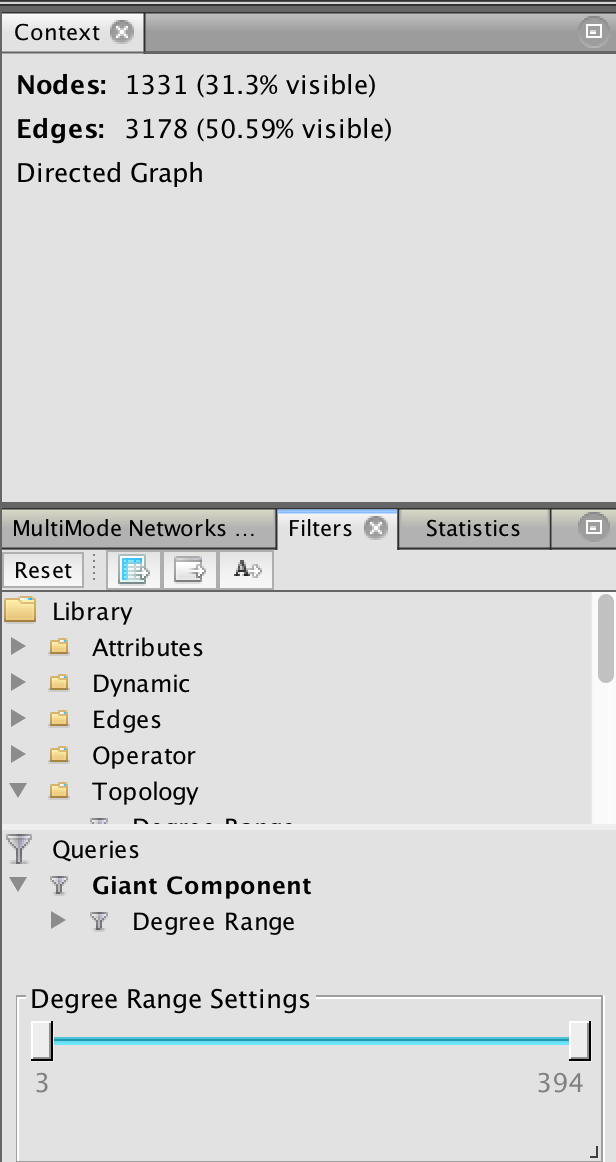
Without using any layout algorithms and doing any filtering, the data looks like a square shaped mess. No useful information can be gleaned from this image other than the fact that there are thousands of nodes and edges connecting these nodes.



The four principal steps in using Gephi are analysis, sizing and coloring, choosing the appropriate layout, and exporting the graph. They’re detailed in the following steps.

## Filtering the Data

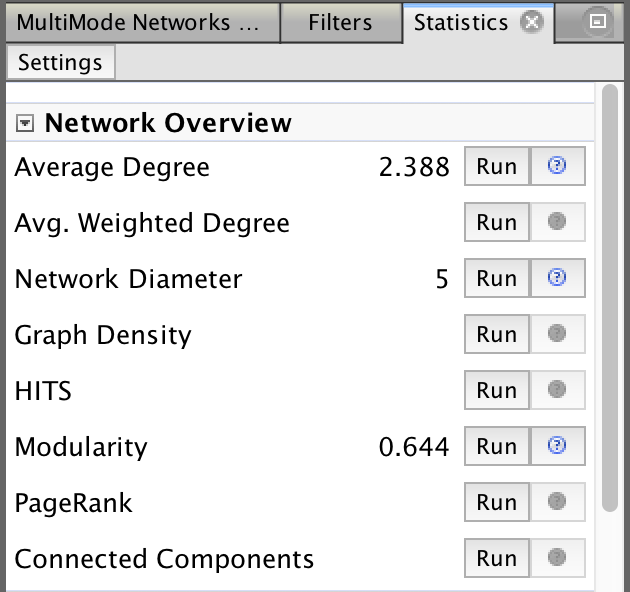
Filtering prunes the graph and keeps only the nodes that are relevant by setting various filtering parameters. Since this data table contained over 4000 nodes and 6000 edges, the Giant Component filter was used to filter out the unconnected nodes. Degree Range filter was used to remove nodes with degrees less than 5. Now the data set is reduced to about 1300 nodes and 3000 edges.



## Getting Statistical Graph Data

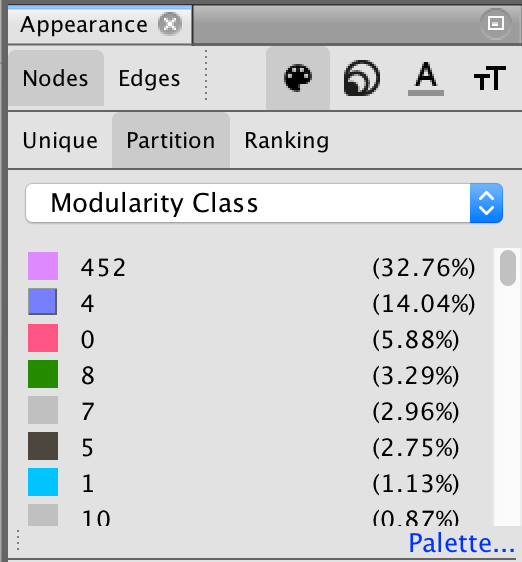
The following algorithms were run and values were computed.

1. Average Degree - The degree of a node in a graph is defined as the number of edges that are incident on that node
2. Network Diameter - It computers three values
   1. Betweenness centrality which measures how often a node appears on shortest paths between nodes in the network
   2. Closeness centrality which is the average distance from a given starting node to all other nodes in the network, and
   3. Eccentricity the distance from a given starting node to the farthest node from it in the network.
3. Modularity - Measures how well a network decomposes into modular communities.
4. PageRank - An iterative algorithm that measures the importance of each node within the network.



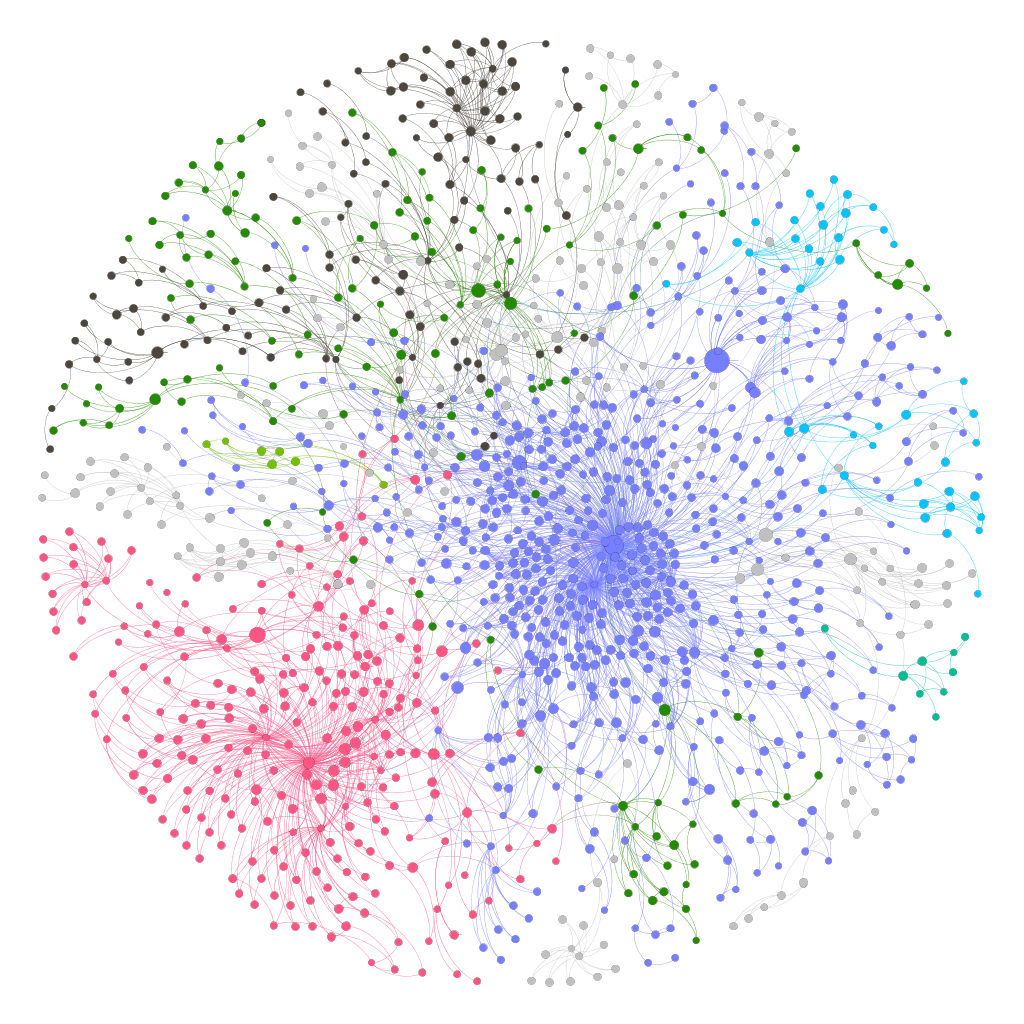
## Appearance

The graph statistics that were computed in the previous step were used to define the color and size of the node and the edge colors. Node colors were chosen based on the Modularity Class.

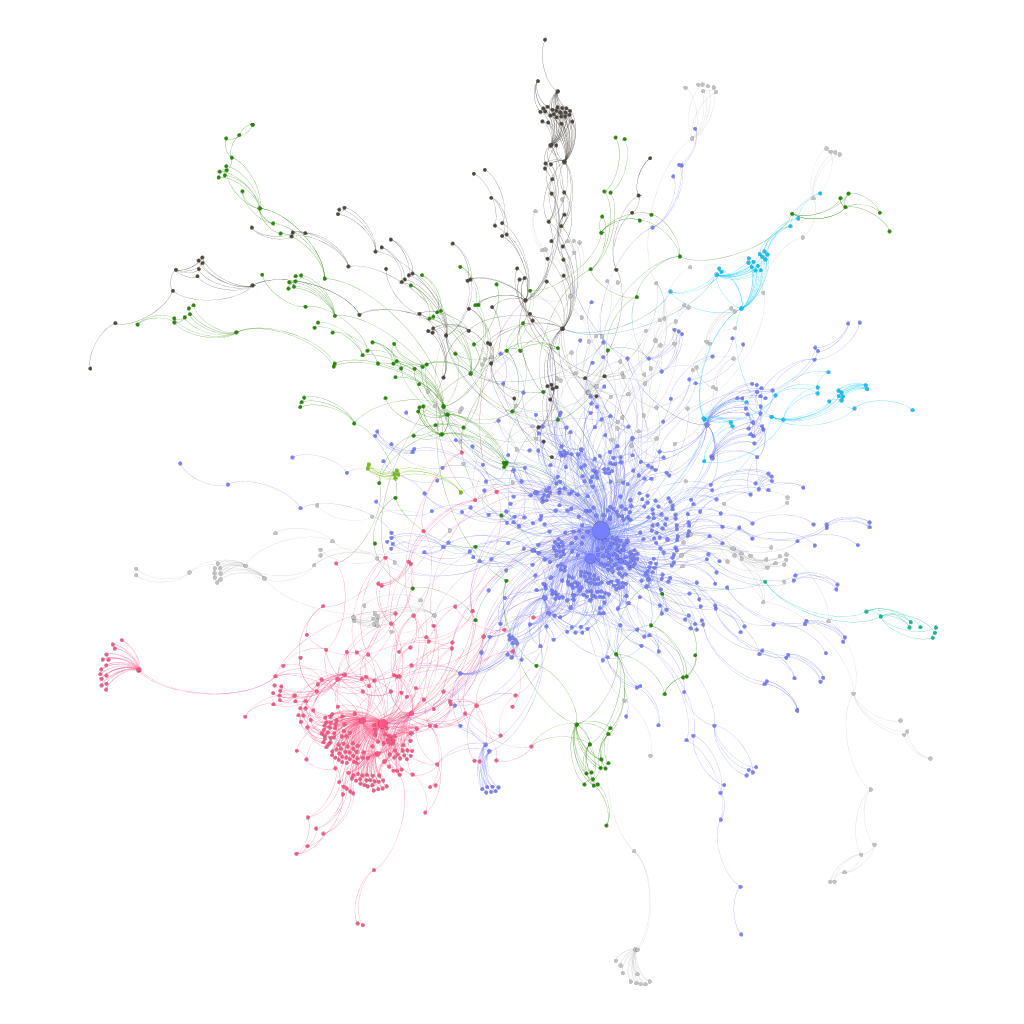


## Layout Algorithms

Force-directed graph drawing algorithms are suitable for visualizing social networks like Twitter. After running the layout algorithm Fruchterman-Reingold, the various communities within the graph are clearly visible. The network colored in blue represents the tweets that were sent or retweeted by @realdonaldtrump or tweets that mentioned him. The pink cluster of data represents tweets that mentioned @johnpodesta which was interesting because Trump mentioned him in his tweet.



When the layout was changed to Yifan Hu, the communities are still visible. While Fruchterman-Reingold showed the graph in a circular form, Yifan Hu layout shows a more disperse graph.



## Exporting the Data

Gephi has built-in tools that can export the visualization in pdf, svg or png format.

## Conclusion

Both Fruchterman-Reingold and Yifan Hu algorithms created aesthetically pleasing visualizations with the communities clearly defined. The Yifan Hu ran noticeably faster and appears less complex.

# Lewis University Facebook Page Network

This dataset was downloaded from the Lewis University Facebook group page and contains different groups and individuals that are connected to Lewis University via the group. The data was generated using the Netvizz app for facebook using ‘Page Like Network’ module. This module starts with a selected page known as the "seed"- Lewis University Facebook page in this example - and retrieves all the pages that page likes. It will continue until the specified crawl depth is reached (2 in this instance). The generated output is a network file in gdf format containing a directed network of pages connected through the likes between them.

## Data Table

The dataset’s Nodes table consists of the node Id, label category, username connected to the group and the kind of activity a user can do. The dataset contains groups varying from news media to non-profit organization and it is interesting to figure out how they interact and who the key components of the network are. Also, contained in the dataset is the Edges table reflecting the relations that exist in this network. This table consisted of the source, target, edge Id, edge type(directed) and weight (all 1).

The two tables are attached to this report.

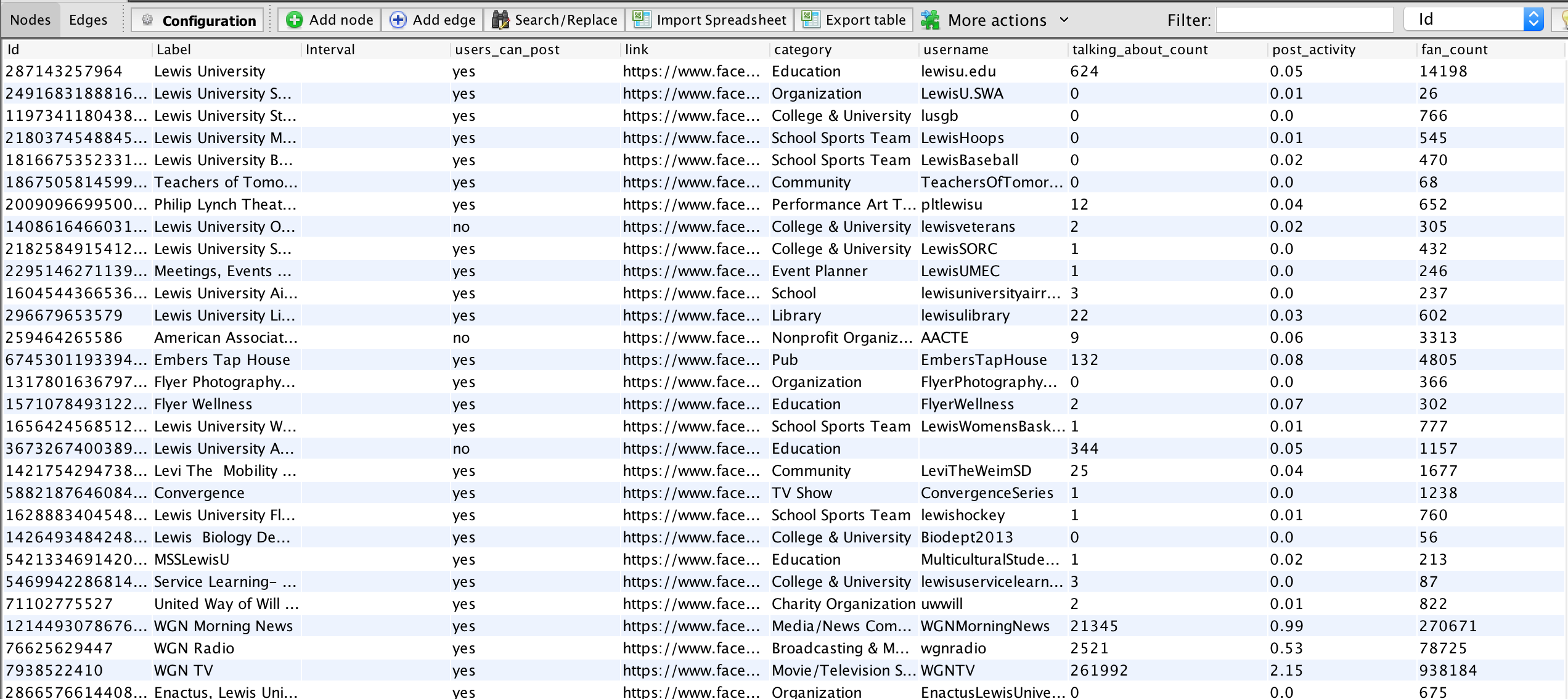


Figure Nodes Table

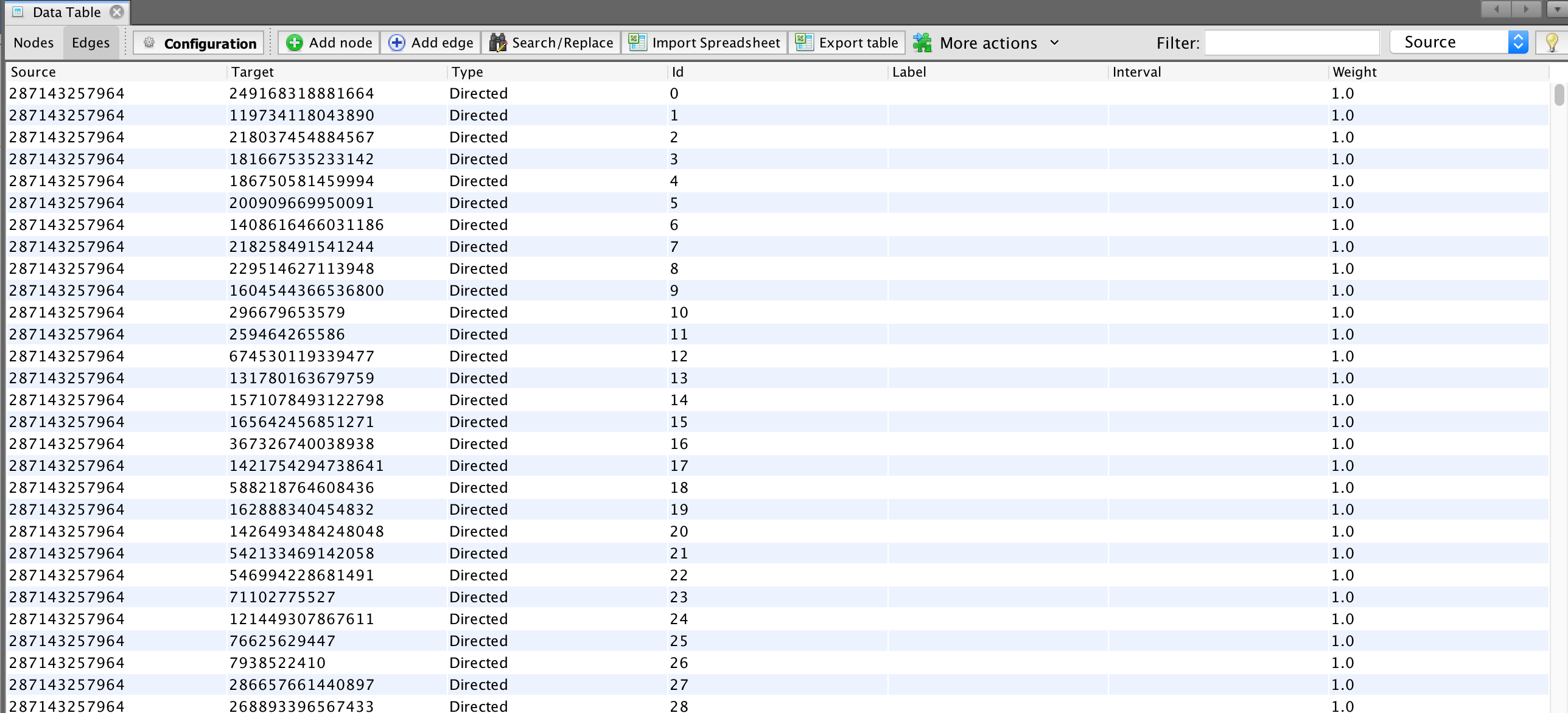
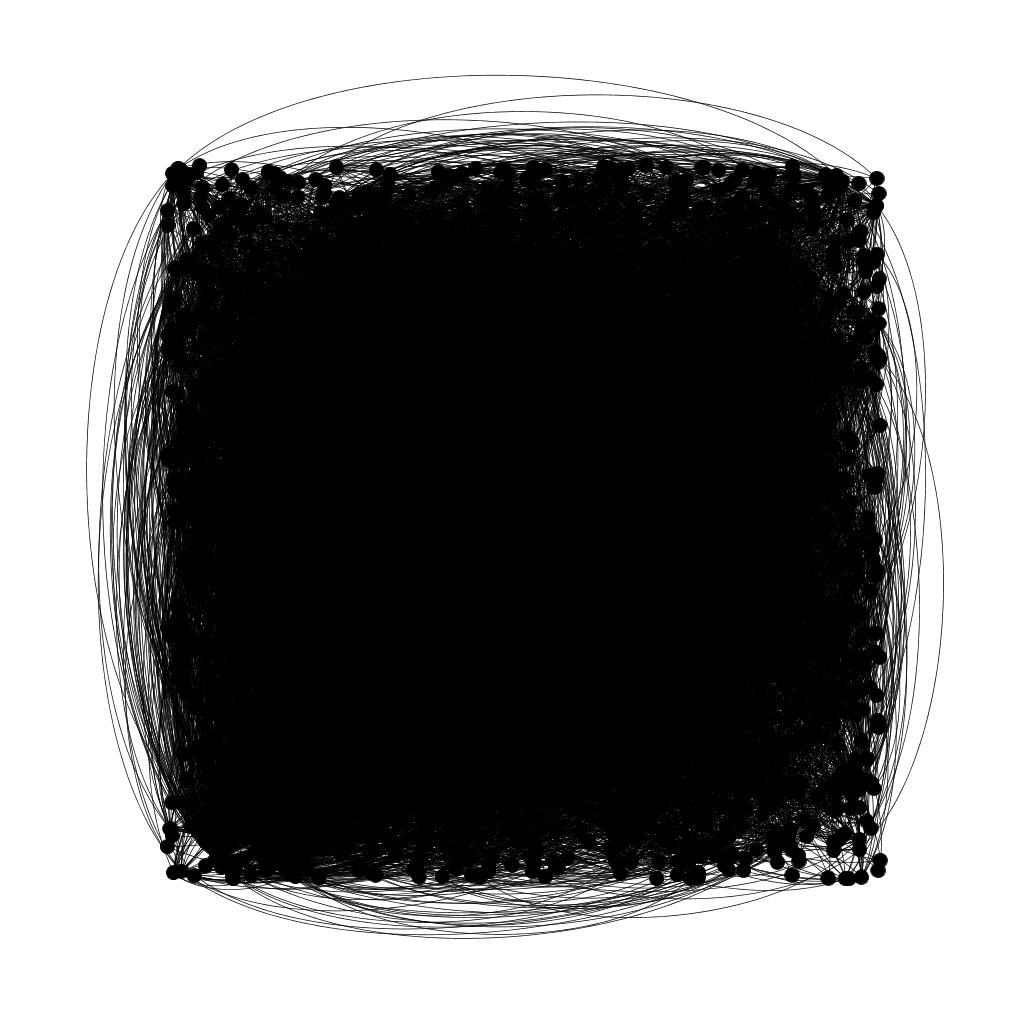


Figure Edges Table

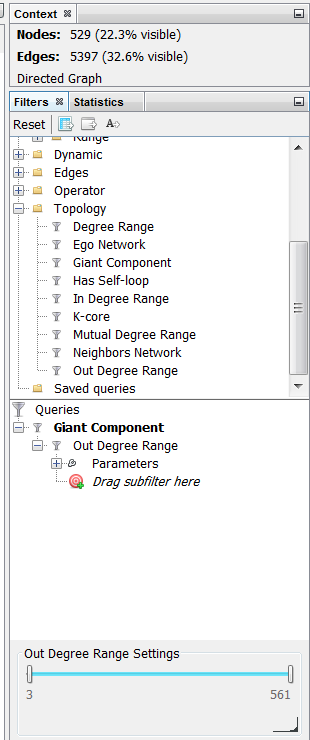
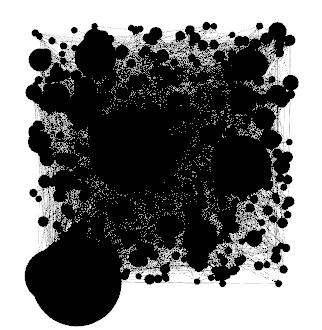
## A First look at the graph

The original graph showed a dense mass of nodes and edges that was difficult to make sense of or to extract any insight from. The graph consisted of 2372 nodes and 16553 edges. To help visualize the original data, the nodes were ranked by size based on their degree of connectivity. Clearly things did improve a little. A few densely condensed nodes could be seen from the graph but offered nothing more. These nodes were considered interesting and warrantied further study.



## Filtering the Data

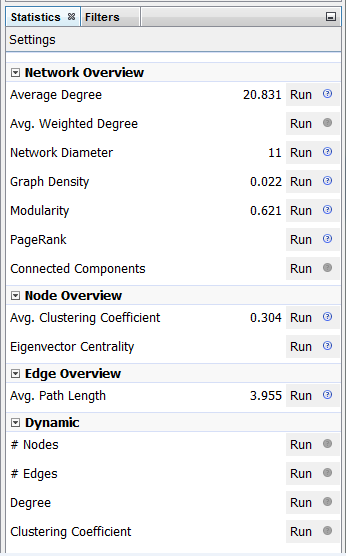
Expanding the graph revealed a lot of nodes lying on the periphery so giant component filtration was applied to the data set to eliminate those loosely associated nodes from the network. However not much was achieved, hence a second type of filtration was necessary. Degree range filtration was applied to the data set to remove nodes with less than. This reduced the nodes to 22.3 % and the edges to 32.6 % of the original values and produced a much clearer and insightful visualization. The visualization reflects two large nodes and a few medium sized notes in addition to many smaller nodes. The large nodes represent the Hispanic Association of Colleges and Universities; and, the American Association of Colleges and Teachers Association. These groups appear prominent in the network. Other important groups include the Bass Schuler Entertainment, and, Joliet Patch as revealed by filtration.

## Getting Statistical Graph

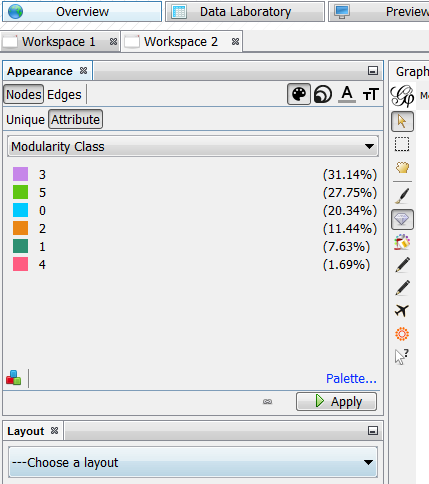
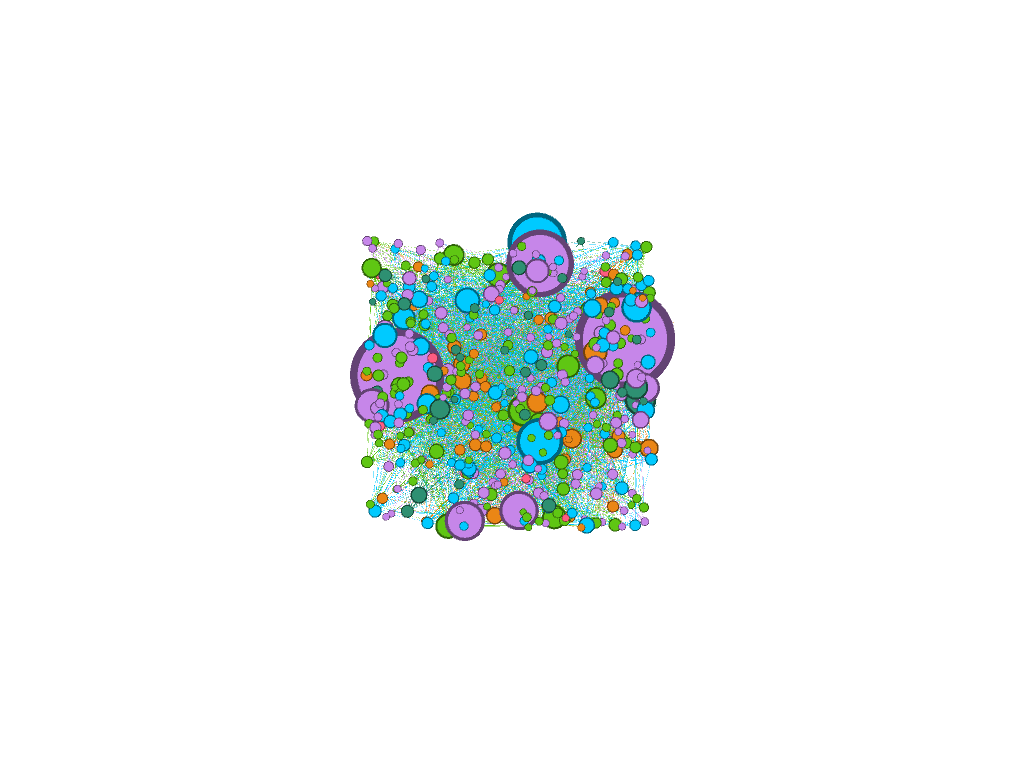
To further enhance the visualization, Statistics of the dataset/graph was explored by playing with degree of measures, density, path length and modularity. This helped identify communities-internal subdivisions in the network. There are methods that permit to highlight these identified communities, which depend on the comparison of the densities of edges within a group, and from the group towards the rest of the network.

The following algorithms were run and values were computed: Average Degree; Network Diameter; Graph Density; Modularity; Average Clustering Coefficient; Average Path Length and PageRank.



## Appearance

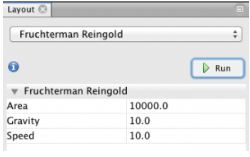
The graph statistics that were computed in the previous step were used to define the color and size of the node and the edge colors. Node colors were chosen based on the Modularity Class.

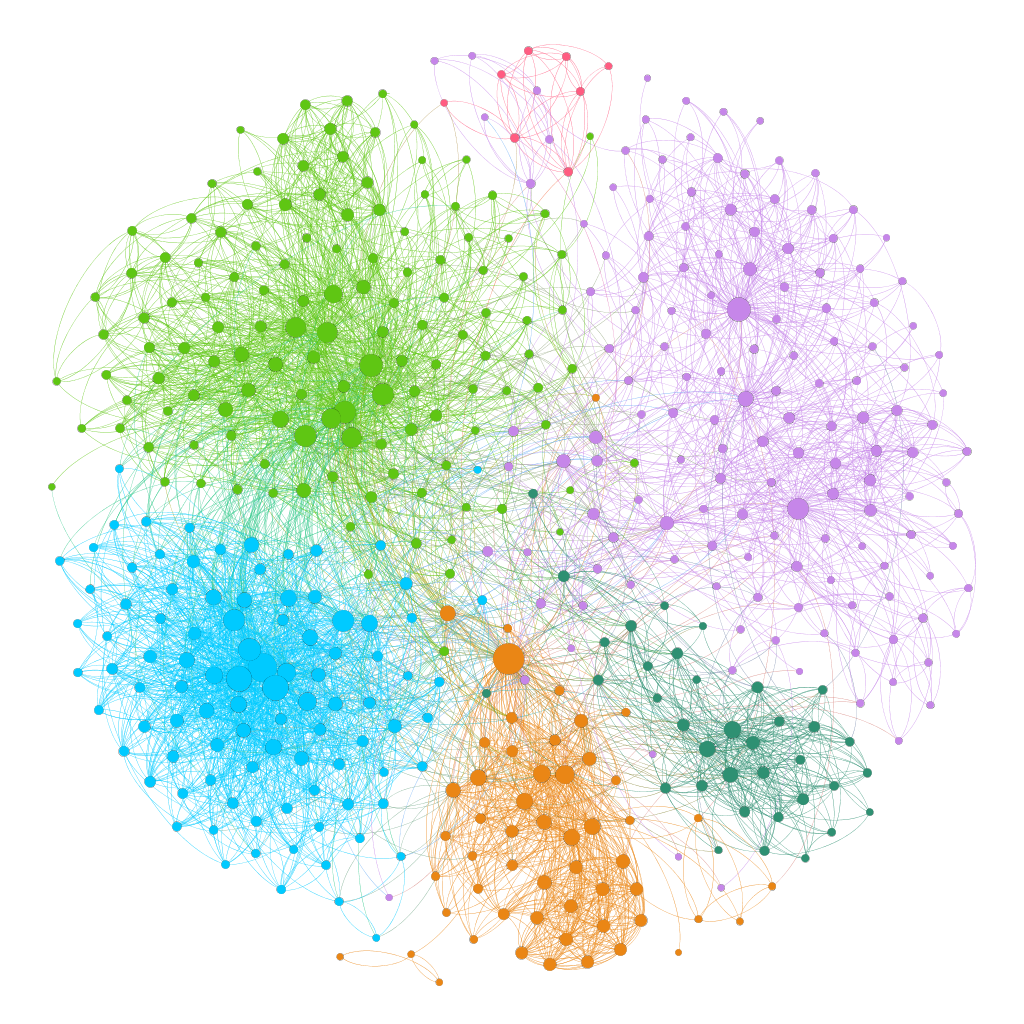
 

Clearly this improved the visualization by color coding the identified communities in the network.

## Layout Algorithms

Initially applied Fruchterman-Reingold layout algorithm. This method disposes nodes based on the strength of the attractive and repulsive powers (gravitational way). Loosely associated nodes are eliminated, leaving distinguishable communities that are densely associated with the network.

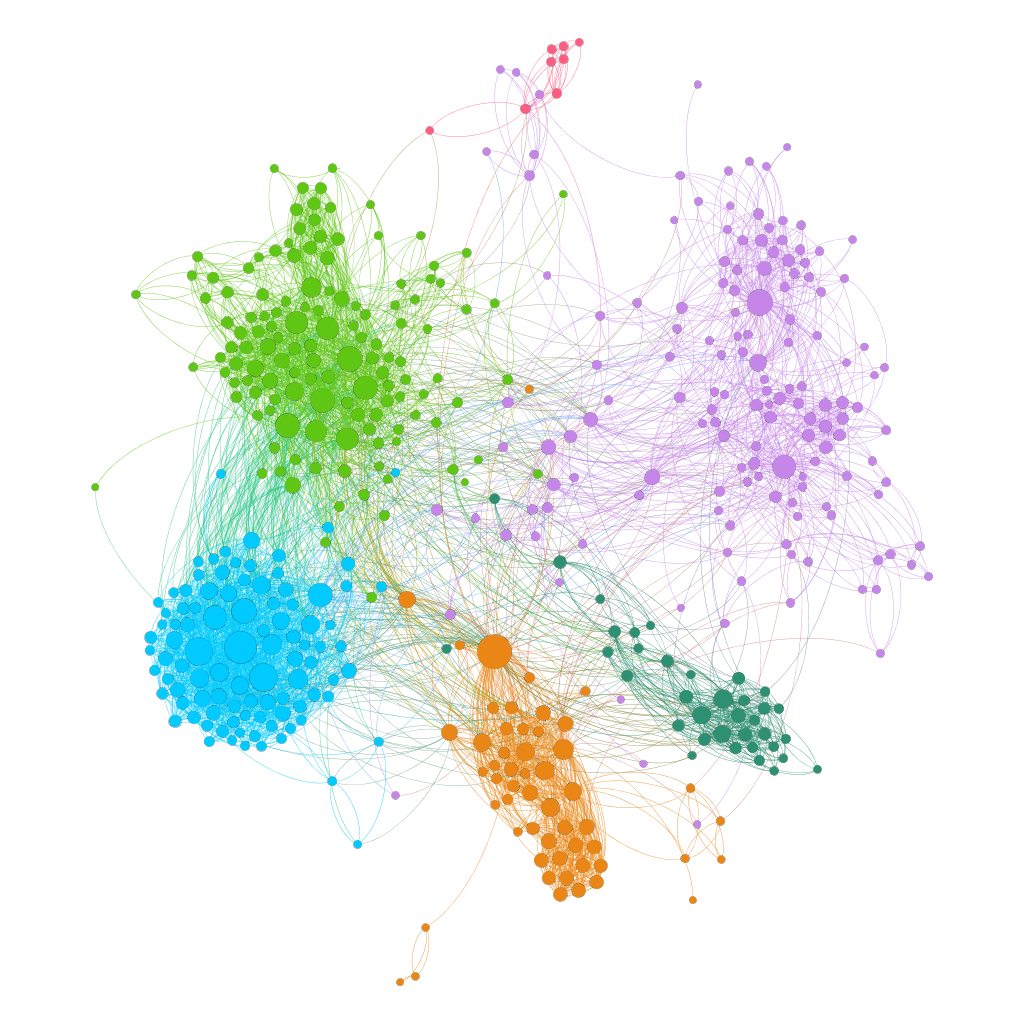




Purple color coding reflects the Hispanic Association of Colleges and Universities; light green coding reflects Joliet Patch; blue coding reflects Chicago Tribune; orange coding reflects Lewis University Recreation Fitness and Wellness; green coding reflects Lasallian link; pink coding reflects Chicago White Socks. American Association of Colleges and Teachers Association is no longer reflected as playing a key role in this network. Chicago Tribune shows strong association with Lewis University Recreation Fitness and Wellness; Hispanic Association of Colleges and Universities; Chicago White Socks. The rest of these groups do not show strong associated between each other. They are loosely associated to one or less other group in the network.

Also applied other layout algorithms to the graph such as ForceAtlas, ForceAtlas2 and Yifan Hu and the results are mainly the same. The only difference is on the dispersion of the graph as shown below. ForceAtlas, ForceAtlas2 and Yifan Hu gave more dispersed results compared to Fruchterman-Reingold layout algorithm.

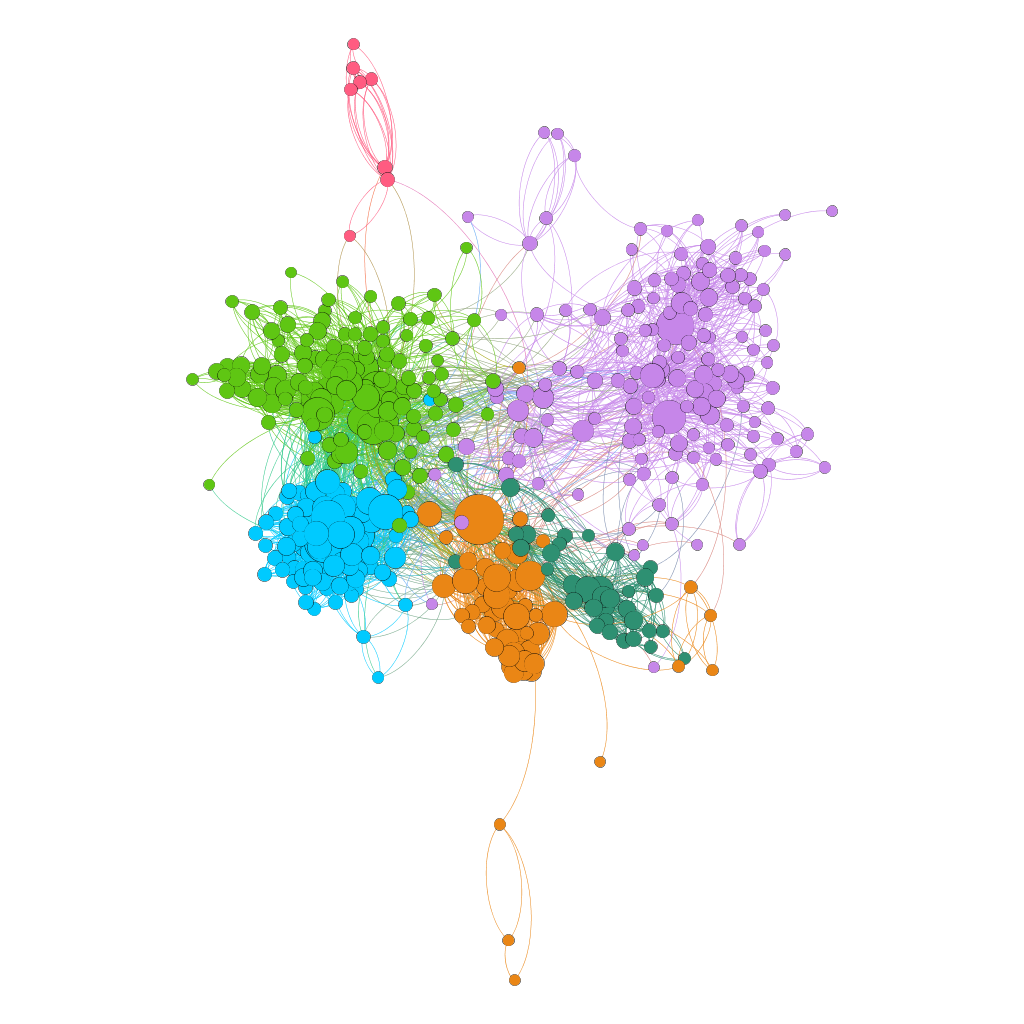
ForceAtlas Layput Algorithm Results:



ForceAtlas Layout Algorithm Results:



Yifan Hu Layout Algorithm Results:

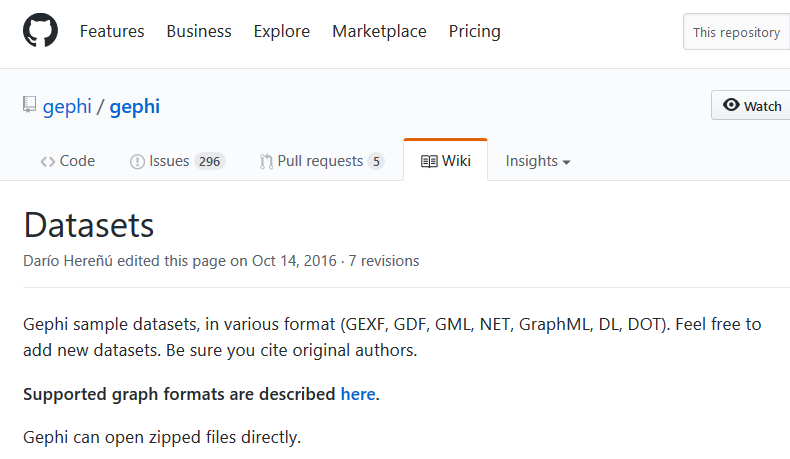


*G*. Conclusion

The Lewis University Facebook group is an association of loosely connected groups with six communities more prominent than the others.

# Jazz Musicians Network

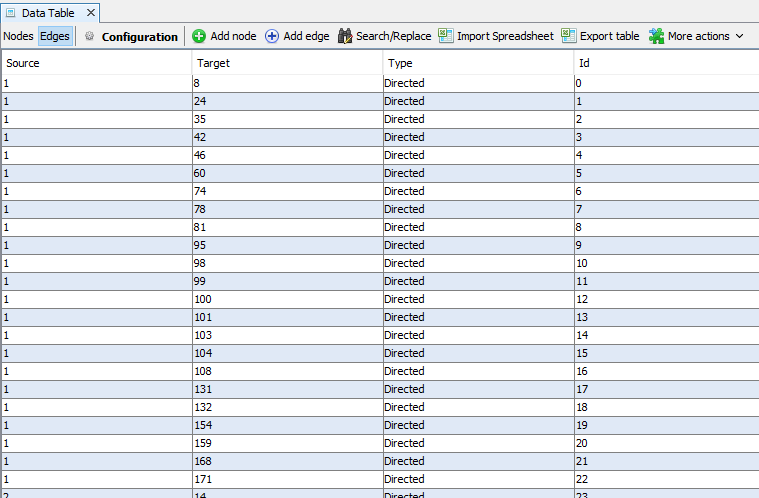
The Jazz Musicians Network dataset was available “Other Networks” Section of the Gephi Github page. <https://github.com/gephi/gephi/wiki/Datasets>. The dataset description is “list of edges of the network of Jazz musicians. The dataset citation list P. Gleiser and L. Danon from the publication Advanced Complex Systems (2003). The data is stored in a .net file format, which loads into Gephi without any error messages.





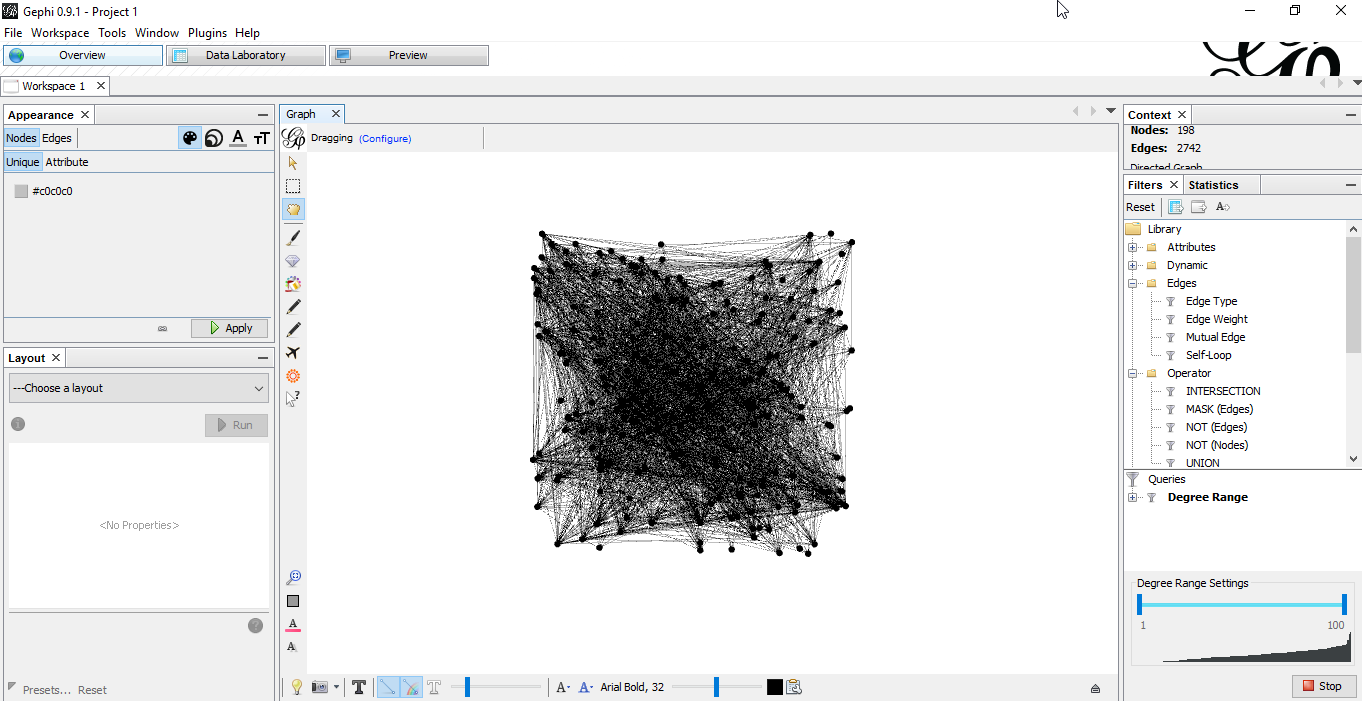
## Data Table

The data table provides the nodes and edges data. The nodes are unlabeled, identified only by an ID number (1 to 198). The edges have an ID, as well as a “source” and “target” ID for the directed edge (0 to 5433).



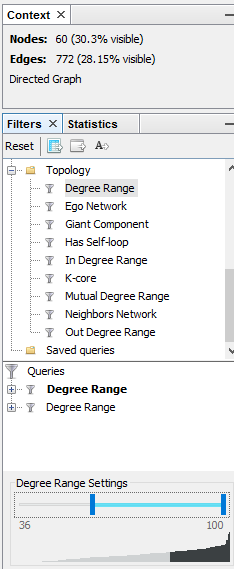
## First Look at the Graph

The Jazz Musicians data is a directed graph made up of 198 nodes and 2742 edges. It’s less dense than other datasets that appear to be a dark solid square. Individual nodes and edges can distinctly be observed, but the center of the graph is very difficult to make sense of. Not much information can be gleaned from this graph without filtering.



## Filtering the Data

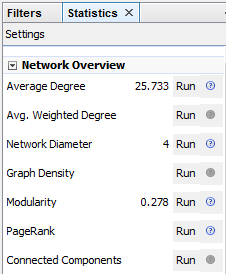
Filtering prunes the graph and keeps only the nodes that are relevant by setting various filtering parameters. Since this data table contained over 198 nodes and 2742 edges, used the degree range filter to remove nodes with degrees less than 36. Now the dataset is reduced to 50 nodes (30.3% visible) and 772 edges (28.15% visible).



## Getting Statistical Graph Data

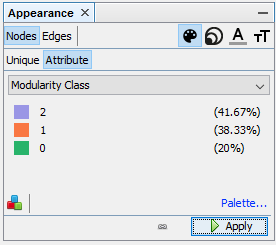
The following algorithms were run and values were computed.

1. Average Degree - The degree of a node in a graph is defined as the number of edges that are incident on that node
2. Network Diameter - It computers three values
   1. Betweenness centrality which measures how often a node appears on shortest paths between nodes in the network
   2. Closeness centrality which is the average distance from a given starting node to all other nodes in the network, and
   3. Eccentricity the distance from a given starting node to the farthest node from it in the network.
3. Modularity - Measures how well a network decomposes into modular communities.
4. PageRank - An iterative algorithm that measures the importance of each node within the network.



## Appearance

The graph statistics that were computed in the previous step were used to define the color and size of the node and the edge colors. Node colors were chosen based on the Modularity Class.

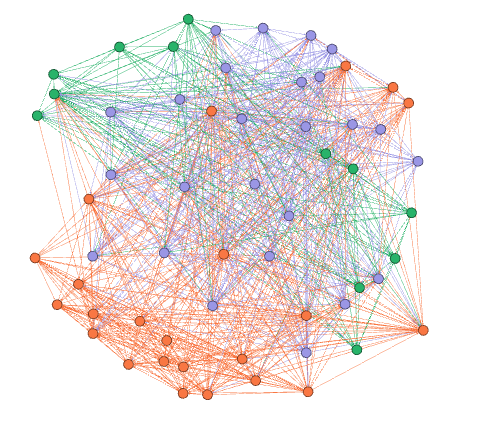


## Layout Algorithms

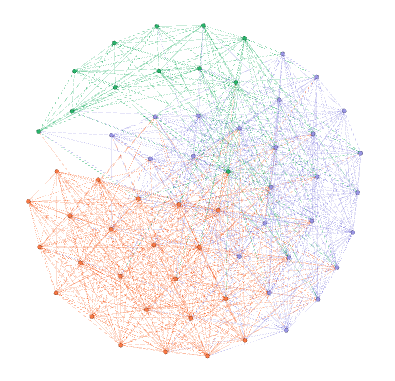
The Noverlap algorithm is a repulsion force to prevent node overlap. The speed was set at 10.0, ratio at 1.2, and the margin at 5.0.



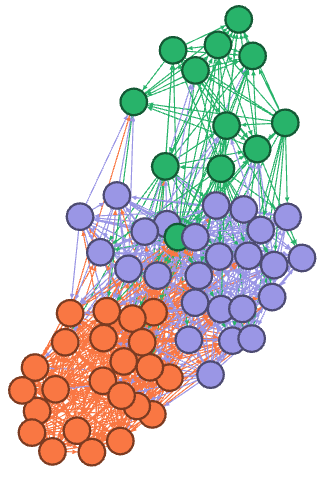
With the Noverlap algorithm



With the Fruchterman Reingold layout algorithm.



With the Yifan Hu layout algorithm.



References

1. https://gephi.org/tutorials/gephi-tutorial-layouts.pdf
2. <https://github.com/seinecle/gephi-tutorials/blob/master/src/main/asciidoc/en/plugins/twitter-streaming-importer-en.adoc>

1. . [↑](#footnote-ref-1)