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**

I. Introduction

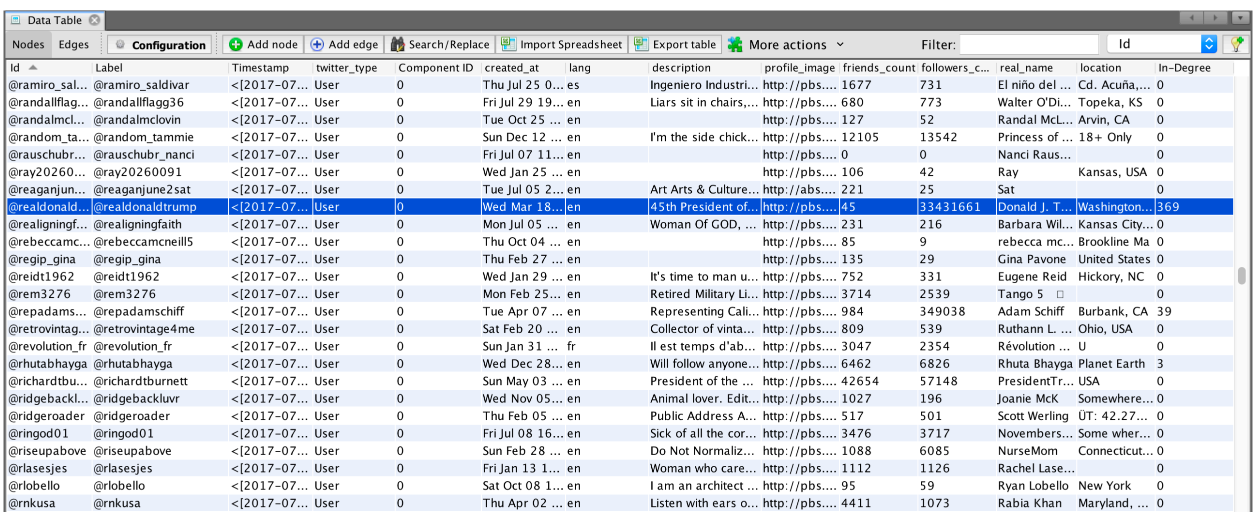
THIS project uses the software Gephi to visualize network data. Gephi has built-in tools to run algorithms like Average Degree, Network Diameter, PageRank, Modularity and generate graph statistic. This data can be analyzed using various layout algorithms such as Fruchterman-Reingold, YifanHu and ForceAtlas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

II. Twitter Network Analysis

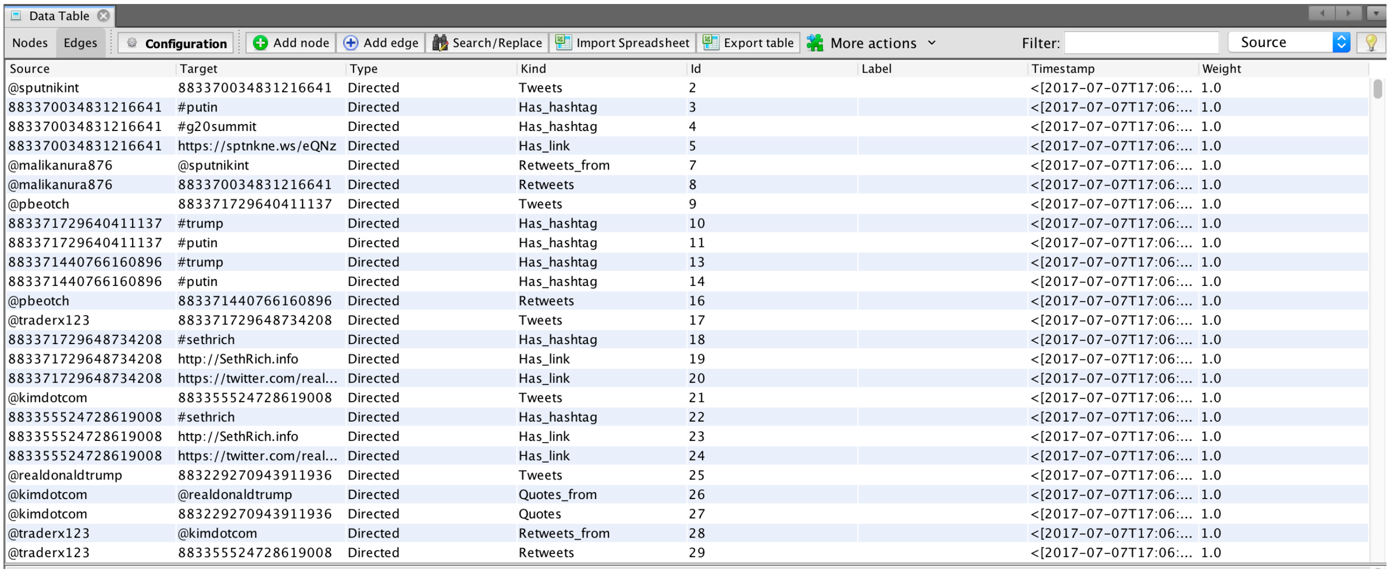
Twitter Streaming Importer is a plug-in that can collect tweets in real-time on the topic of your choice. It will generate a data table and create connections between the users mentioned in those tweets. Since the G-20 summit is happening right now and Donald Trump is a prolific Twitter user, I decided 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.

A. 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.



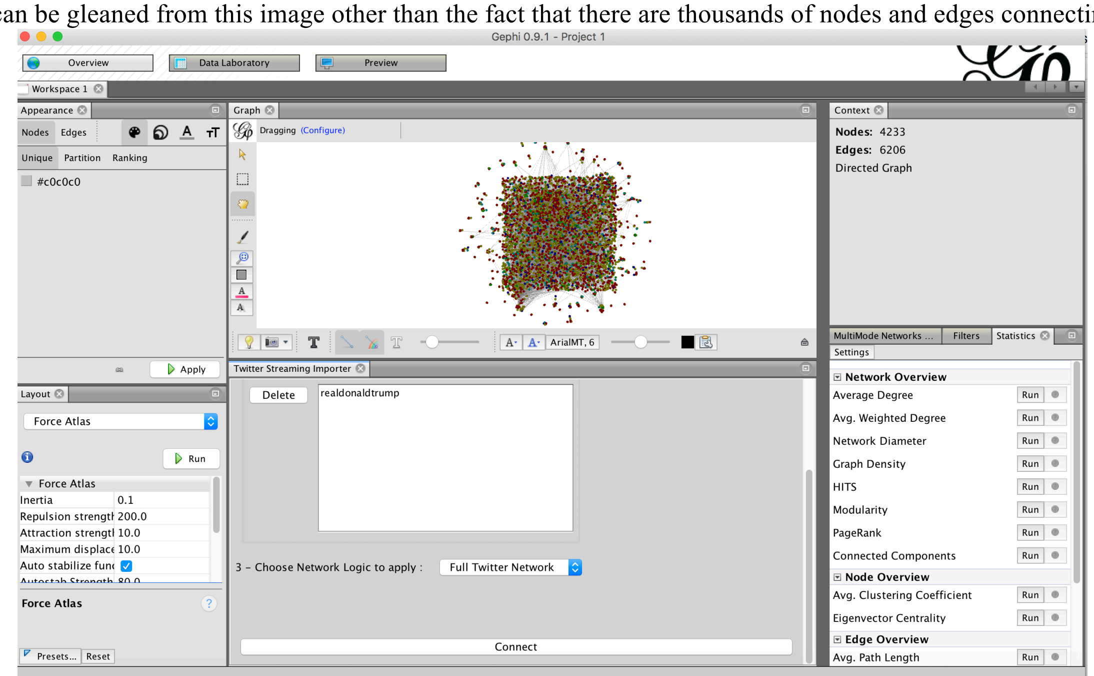
Id	Label	Timestamp	twitter_type	Component ID	created_at	lang	description	profile_image	friends_count	followers_c...	real_name	location	In-Degree
@ramiro_sal...	@ramiro_saldivar	<[2017-07...	User	0	Thu Jul 25 0...	es	Ingeniero Industri...	http://pbs....	1677	731	El niño del ...	Cd. Acuña,...	0
@randallflag...	@randallflagg36	<[2017-07...	User	0	Fri Jul 29 19...	en	Liars sit in chairs,...	http://pbs....	680	773	Walter O'Di...	Topeka, KS	0
@randalmcl...	@randalmclovin	<[2017-07...	User	0	Tue Oct 25 ...	en		http://pbs....	127	52	Randal McL...	Arvin, CA	0
@random_ta...	@random_tammie	<[2017-07...	User	0	Sun Dec 12 ...	en	I'm the side chick...	http://pbs....	12105	13542	Princess of ...	18+ Only	0
@rauschubr...	@rauschubr_nanci	<[2017-07...	User	0	Fri Jul 07 11...	en		http://pbs....	0	0	Nanci Raus...		0
@ray20260...	@ray20260091	<[2017-07...	User	0	Wed Jan 25 ...	en		http://pbs....	106	42	Ray	Kansas, USA	0
@reaganjun...	@reaganjune2sat	<[2017-07...	User	0	Tue Jul 05 2...	en	Art Arts & Culture...	http://abs....	221	25	Sat		0
@realdonald...	@realdonaldtrump	<[2017-07...	User	0	Wed Mar 18...	en	45th President of...	http://pbs....	45	33431661	Donald J. T...	Washington...	369
@realigningf...	@realigningfaith	<[2017-07...	User	0	Mon Jul 05 ...	en	Woman Of GOD, ...	http://pbs....	231	216	Barbara Wil...	Kansas City...	0
@rebeccamc...	@rebeccamcneill5	<[2017-07...	User	0	Thu Oct 04 ...	en		http://pbs....	85	9	rebecca mc...	Brookline Ma	0
@regip_gina	@regip_gina	<[2017-07...	User	0	Thu Feb 27 ...	en		http://pbs....	135	29	Gina Pavone	United States	0
@reidt1962	@reidt1962	<[2017-07...	User	0	Wed Jan 29 ...	en	It's time to man u...	http://pbs....	752	331	Eugene Reid	Hickory, NC	0
@rem3276	@rem3276	<[2017-07...	User	0	Mon Feb 25...	en	Retired Military Li...	http://pbs....	3714	2539	Tango 5 □		0
@repadams...	@repadamschiff	<[2017-07...	User	0	Tue Apr 07 ...	en	Representing Cali...	http://pbs....	984	349038	Adam Schiff	Burbank, CA	39
@retrovintag...	@retrovintage4me	<[2017-07...	User	0	Sat Feb 20 ...	en	Collector of vinta...	http://pbs....	809	539	Ruthann L. ...	Ohio, USA	0
@revolution_fr	@revolution_fr	<[2017-07...	User	0	Sun Jan 31 ...	fr	Il est temps d'ab...	http://pbs....	3047	2354	Révolution ...	U	0
@rhutabhayga	@rhutabhayga	<[2017-07...	User	0	Wed Dec 28...	en	Will follow anyone...	http://pbs....	6462	6826	Rhuta Bhayga	Planet Earth	3
@richardtbu...	@richardtburnett	<[2017-07...	User	0	Sun May 03 ...	en	President of the ...	http://pbs....	42654	57148	PresidentTr...	USA	0
@ridgebackl...	@ridgebackluvr	<[2017-07...	User	0	Wed Nov 05...	en	Animal lover. Edit...	http://pbs....	1027	196	Joanie McK	Somewhere...	0
@ridgeroader	@ridgeroader	<[2017-07...	User	0	Thu Feb 05 ...	en	Public Address A...	http://pbs....	517	501	Scott Werling	ÜT: 42.27...	0
@ringod01	@ringod01	<[2017-07...	User	0	Fri Jul 08 16...	en	Sick of all the cor...	http://pbs....	3476	3717	Novembers...	Some wher...	0
@riseupabove	@riseupabove	<[2017-07...	User	0	Sun Feb 28 ...	en	Do Not Normaliz...	http://pbs....	1088	6085	NurseMom	Connecticut...	0
@rlasesjes	@rlasesjes	<[2017-07...	User	0	Fri Jan 13 1...	en	Woman who care...	http://pbs....	1112	1126	Rachel Lase...		0
@rlobello	@rlobello	<[2017-07...	User	0	Sat Oct 08 1...	en	I am an architect ...	http://pbs....	95	59	Ryan Lobello	New York	0
@rnkusa	@rnkusa	<[2017-07...	User	0	Thu Apr 02 ...	en	Listen with ears o...	http://pbs....	4411	1073	Rabia Khan	Maryland, ...	0



Source	Target	Type	Kind	Id	Label	Timestamp	Weight
@sputnikint	883370034831216641	Directed	Tweets	2		<[2017-07-07T17:06:...	1.0
883370034831216641	#putin	Directed	Has_hashtag	3		<[2017-07-07T17:06:...	1.0
883370034831216641	#g20summit	Directed	Has_hashtag	4		<[2017-07-07T17:06:...	1.0
883370034831216641	https://sptnkne.ws/eQNz	Directed	Has_link	5		<[2017-07-07T17:06:...	1.0
@malikanura876	@sputnikint	Directed	Retweets_from	7		<[2017-07-07T17:06:...	1.0
@malikanura876	883370034831216641	Directed	Retweets	8		<[2017-07-07T17:06:...	1.0
@pbeotch	883371729640411137	Directed	Tweets	9		<[2017-07-07T17:06:...	1.0
883371729640411137	#trump	Directed	Has_hashtag	10		<[2017-07-07T17:06:...	1.0
883371729640411137	#putin	Directed	Has_hashtag	11		<[2017-07-07T17:06:...	1.0
883371440766160896	#trump	Directed	Has_hashtag	13		<[2017-07-07T17:06:...	1.0
883371440766160896	#putin	Directed	Has_hashtag	14		<[2017-07-07T17:06:...	1.0
@pbeotch	883371440766160896	Directed	Retweets	16		<[2017-07-07T17:06:...	1.0
@traderx123	883371729648734208	Directed	Tweets	17		<[2017-07-07T17:06:...	1.0
883371729648734208	#sethrich	Directed	Has_hashtag	18		<[2017-07-07T17:06:...	1.0
883371729648734208	http://SethRich.info	Directed	Has_link	19		<[2017-07-07T17:06:...	1.0
883371729648734208	https://twitter.com/real...	Directed	Has_link	20		<[2017-07-07T17:06:...	1.0
@kimdotcom	883355524728619008	Directed	Tweets	21		<[2017-07-07T17:06:...	1.0
883355524728619008	#sethrich	Directed	Has_hashtag	22		<[2017-07-07T17:06:...	1.0
883355524728619008	http://SethRich.info	Directed	Has_link	23		<[2017-07-07T17:06:...	1.0
883355524728619008	https://twitter.com/real...	Directed	Has_link	24		<[2017-07-07T17:06:...	1.0
@realdonaldtrump	883229270943911936	Directed	Tweets	25		<[2017-07-07T17:06:...	1.0
@kimdotcom	@realdonaldtrump	Directed	Quotes_from	26		<[2017-07-07T17:06:...	1.0
@kimdotcom	883229270943911936	Directed	Quotes	27		<[2017-07-07T17:06:...	1.0
@traderx123	@kimdotcom	Directed	Retweets_from	28		<[2017-07-07T17:06:...	1.0
@traderx123	883355524728619008	Directed	Retweets	29		<[2017-07-07T17:06:...	1.0

B. A First Look at the Graph

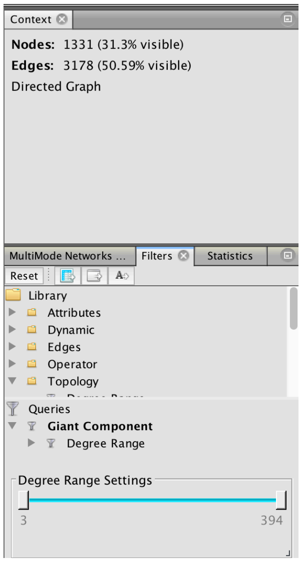
Without using any layout algorithms and doing any filtering, the data looks like a square shaped mess. No useful information can be gleaned from this image other than the fact that there are thousands of nodes and edges connecting these nodes.



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.

C. Filtering the Data

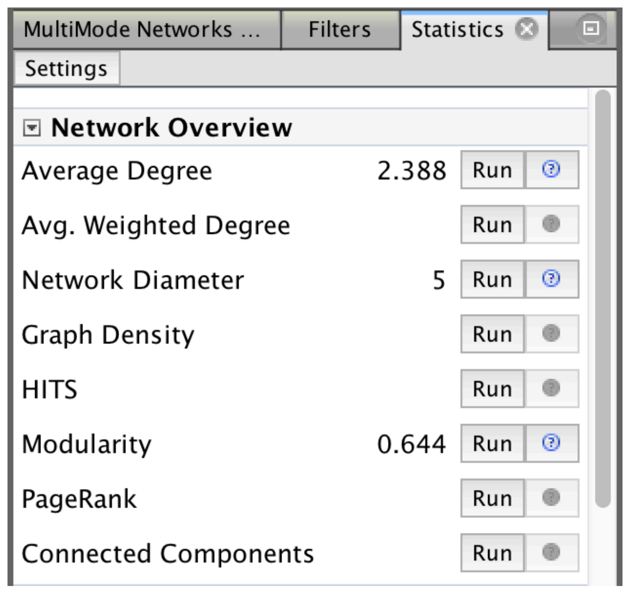
Filtering prunes the graph and keeps only the nodes that are relevant by setting various filtering parameters. Since this data table contained over 4000 nodes and 6000 edges, the Giant Component filter was used to filter out the unconnected nodes. Degree Range filter was used to remove nodes with degrees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



D. Getting Statistical Graph Data

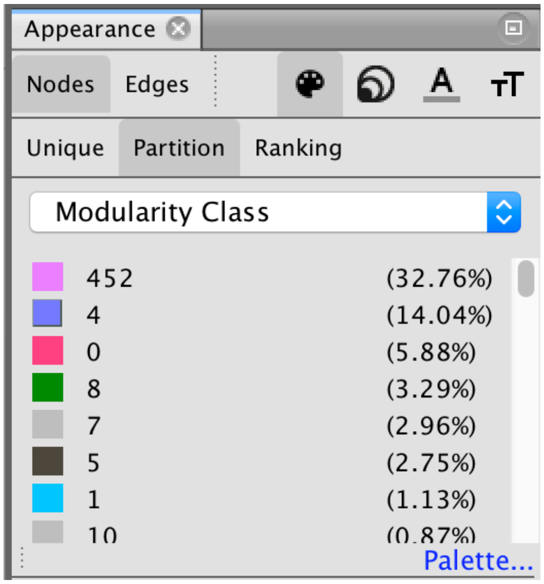
The following algorithms were run and values were computed.

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
   a. Betweenness centrality which measures how often a node appears on shortest paths between nodes in the network
   b. Closeness centrality which is the average distance from a given starting node to all other nodes in the network, and
   c. 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.



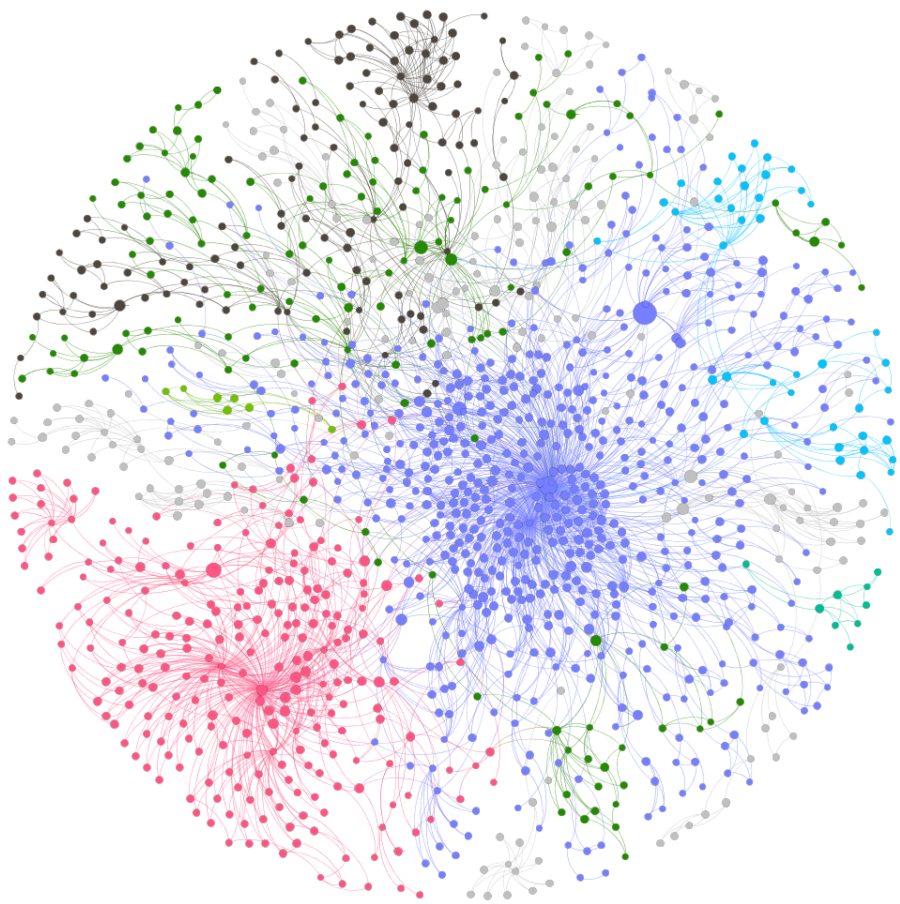
E. Appearance

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.

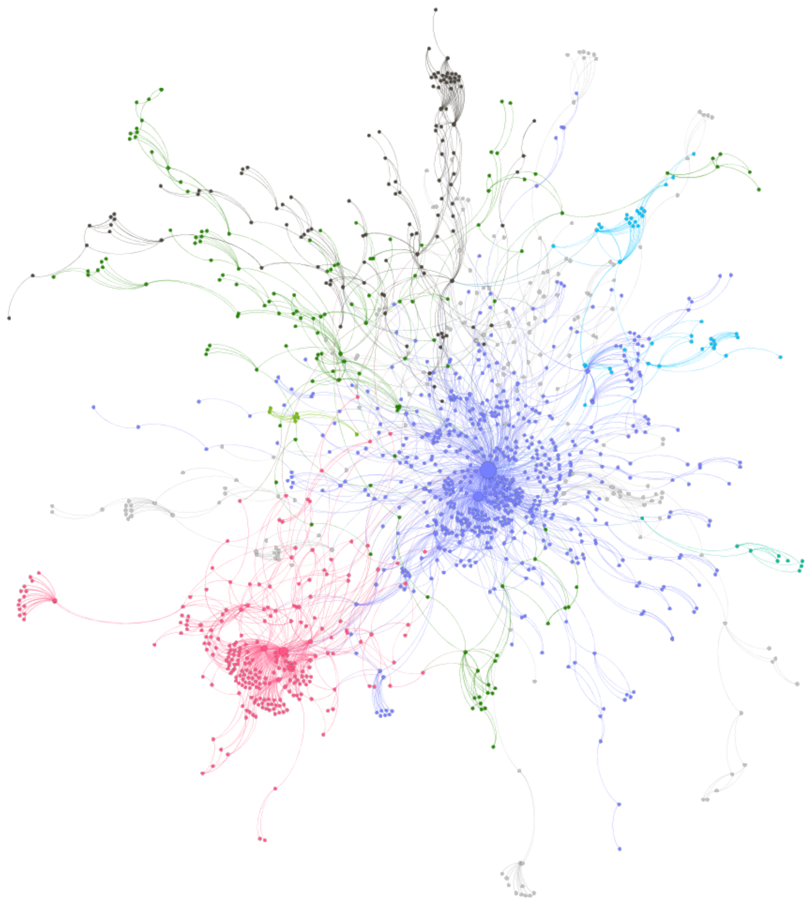


F. Layout Algorithms

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.



G. Exporting the Data

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

H. 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.

III. 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.

A. 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.

Nodes Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More actions | Filter: | Id

Id	Label	Interval	users_can_post	link	category	username	talking_about_count	post_activity	fan_count
287143257964	Lewis University		yes	https://www.face...	Education	lewisu.edu	624	0.05	14198
2491683188816...	Lewis University S...		yes	https://www.face...	Organization	LewisU.SWA	0	0.01	26
1197341180438...	Lewis University St...		yes	https://www.face...	College & University	lusgb	0	0.0	766
2180374548845...	Lewis University M...		yes	https://www.face...	School Sports Team	LewisHoops	0	0.01	545
1816675352331...	Lewis University B...		yes	https://www.face...	School Sports Team	LewisBaseball	0	0.02	470
1867505814599...	Teachers of Tomo...		yes	https://www.face...	Community	TeachersOfTomor...	0	0.0	68
2009096699500...	Philip Lynch Theat...		yes	https://www.face...	Performance Art T...	pltlewisu	12	0.04	652
1408616466031...	Lewis University O...		no	https://www.face...	College & University	lewisveterans	2	0.02	305
2182584915412...	Lewis University S...		yes	https://www.face...	College & University	LewisSORC	1	0.0	432
2295146271139...	Meetings, Events ...		yes	https://www.face...	Event Planner	LewisUMEC	1	0.0	246
1604544366536...	Lewis University Ai...		yes	https://www.face...	School	lewisuniversityairr...	3	0.0	237
296679653579	Lewis University Li...		yes	https://www.face...	Library	lewisulibrary	22	0.03	602
259464265586	American Associat...		no	https://www.face...	Nonprofit Organiz...	AACTE	9	0.06	3313
6745301193394...	Embers Tap House		yes	https://www.face...	Pub	EmbersTapHouse	132	0.08	4805
1317801636797...	Flyer Photography...		yes	https://www.face...	Organization	FlyerPhotography...	0	0.0	366
1571078493122...	Flyer Wellness		yes	https://www.face...	Education	FlyerWellness	2	0.07	302
1656424568512...	Lewis University W...		yes	https://www.face...	School Sports Team	LewisWomensBask...	1	0.01	777
3673267400389...	Lewis University A...		no	https://www.face...	Education		344	0.05	1157
1421754294738...	Levi The Mobility ...		yes	https://www.face...	Community	LeviTheWeimSD	25	0.04	1677
5882187646084...	Convergence		yes	https://www.face...	TV Show	ConvergenceSeries	1	0.0	1238
1628883404548...	Lewis University Fl...		yes	https://www.face...	School Sports Team	lewishockey	1	0.01	760
1426493484248...	Lewis Biology De...		yes	https://www.face...	College & University	Biodept2013	0	0.0	56
5421334691420...	MSSLewisU		yes	https://www.face...	Education	MulticulturalStude...	1	0.02	213
5469942286814...	Service Learning- ...		yes	https://www.face...	College & University	lewisuservicelearn...	3	0.0	87
71102775527	United Way of Will ...		yes	https://www.face...	Charity Organization	uwwill	2	0.01	822
1214493078676...	WGN Morning News		yes	https://www.face...	Media/News Com...	WGNMorningNews	21345	0.99	270671
76625629447	WGN Radio		yes	https://www.face...	Broadcasting & M...	wgnradio	2521	0.53	78725
7938522410	WGN TV		yes	https://www.face...	Movie/Television S...	WGNTV	261992	2.15	938184
2866576614408...	Enactus, Lewis Uni...		yes	https://www.face...	Organization	EnactusLewisUnive...	0	0.0	675

Figure 1 Nodes Table

Data Table

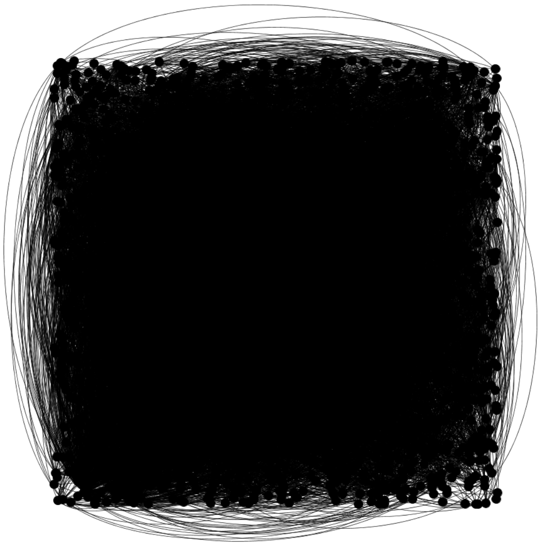
Nodes Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More actions | Filter: | Source

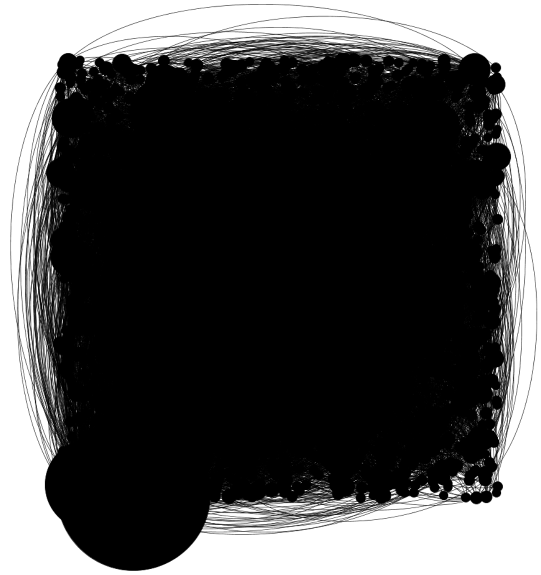
Source	Target	Type	Id	Label	Interval	Weight
287143257964	249168318881664	Directed	0			1.0
287143257964	119734118043890	Directed	1			1.0
287143257964	218037454884567	Directed	2			1.0
287143257964	181667535233142	Directed	3			1.0
287143257964	186750581459994	Directed	4			1.0
287143257964	200909669950091	Directed	5			1.0
287143257964	1408616466031186	Directed	6			1.0
287143257964	218258491541244	Directed	7			1.0
287143257964	229514627113948	Directed	8			1.0
287143257964	1604544366536800	Directed	9			1.0
287143257964	296679653579	Directed	10			1.0
287143257964	259464265586	Directed	11			1.0
287143257964	674530119339477	Directed	12			1.0
287143257964	131780163679759	Directed	13			1.0
287143257964	1571078493122798	Directed	14			1.0
287143257964	165642456851271	Directed	15			1.0
287143257964	367326740038938	Directed	16			1.0
287143257964	1421754294738641	Directed	17			1.0
287143257964	588218764608436	Directed	18			1.0
287143257964	162888340454832	Directed	19			1.0
287143257964	1426493484248048	Directed	20			1.0
287143257964	542133469142058	Directed	21			1.0
287143257964	546994228681491	Directed	22			1.0
287143257964	71102775527	Directed	23			1.0
287143257964	121449307867611	Directed	24			1.0
287143257964	76625629447	Directed	25			1.0
287143257964	7938522410	Directed	26			1.0
287143257964	286657661440897	Directed	27			1.0
287143257964	268893396567433	Directed	28			1.0

Figure 2 Edges Table

B. A First look at the graph

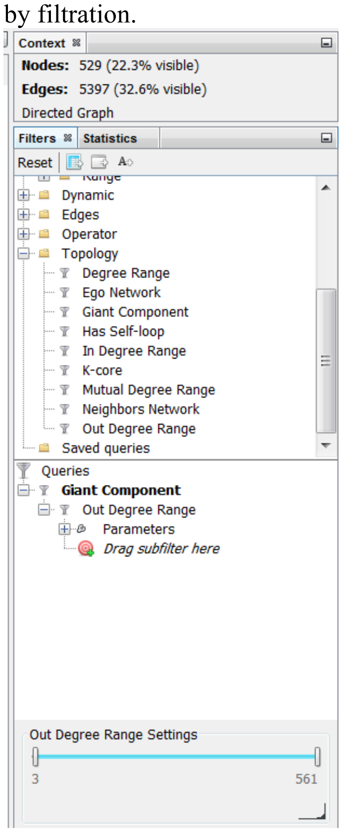
The original graph showed a dense mass of nodes and edges that was difficult to make sense of or to extract any insight from. The graph consisted of 2372 nodes and 16553 edges. To help visualize the original data, the nodes were ranked by size based on their degree of connectivity. Clearly things did 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.

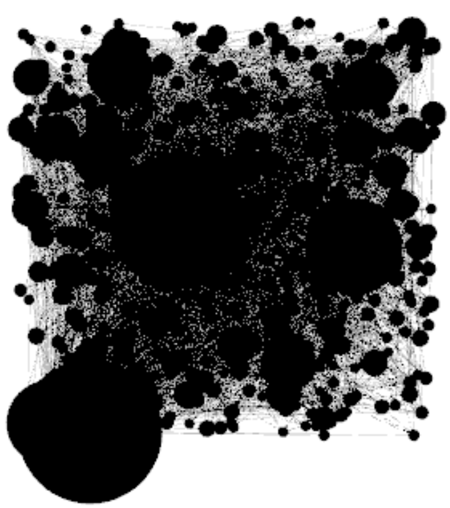




C. Filtering the Data

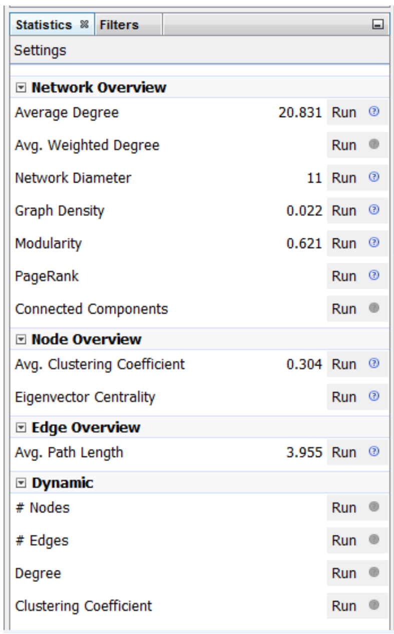
Expanding the graph revealed a lot of nodes lying on the periphery so giant component filtration was applied to the data set to eliminate those loosely associated nodes from the network. However not much was achieved, hence a second type of filtration was necessary. Degree range filtration was applied 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.





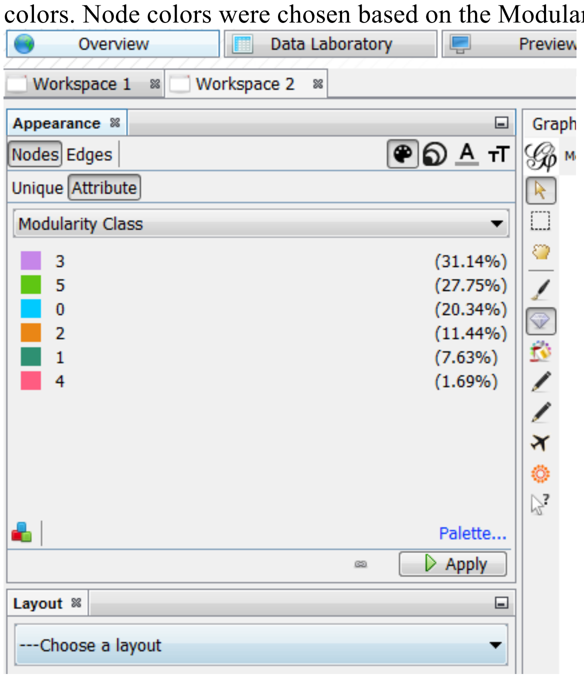
D. Getting Statistical Graph

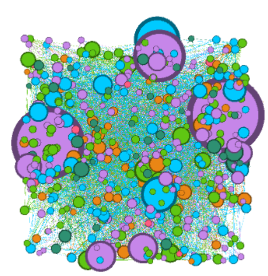
To further enhance the visualization, Statistics of the dataset/graph was explored by playing with degree of measures, density, path length and modularity. This helped identify communities-internal subdivisions in the network. There are methods that permit to highlight these identified communities, which 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.



E. 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.

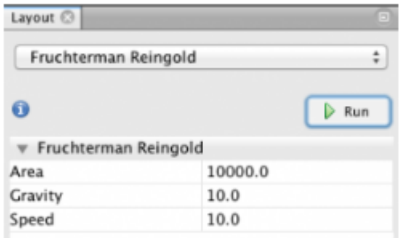


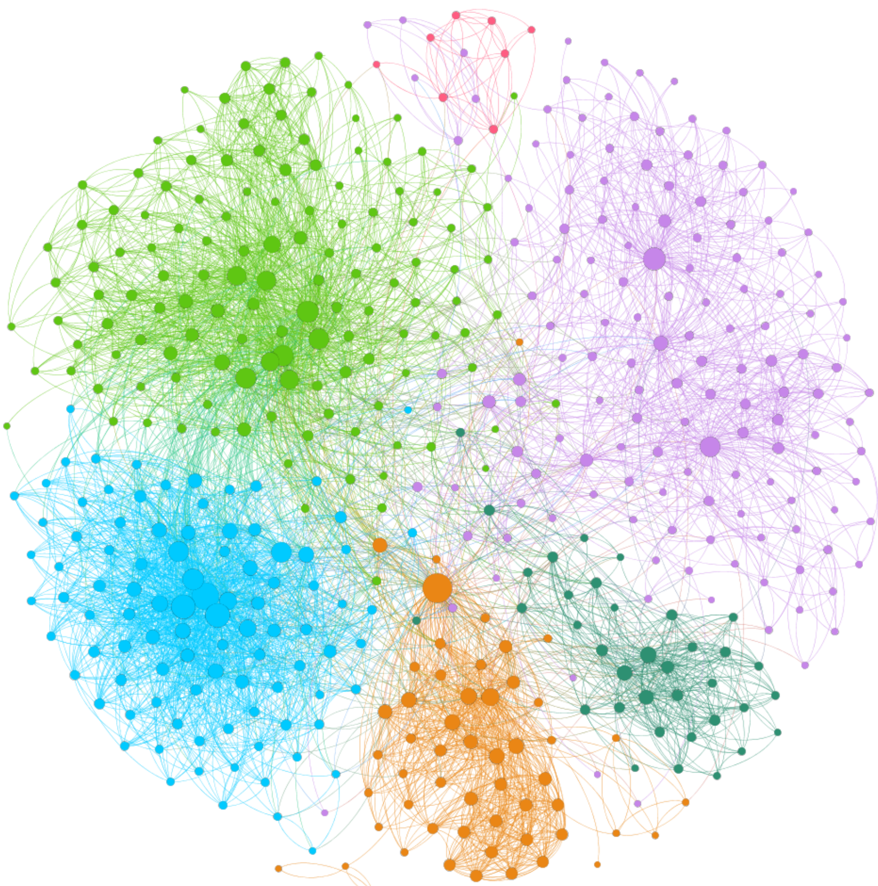


Clearly this improved the visualization by color coding the identified communities in the network.

F. 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.

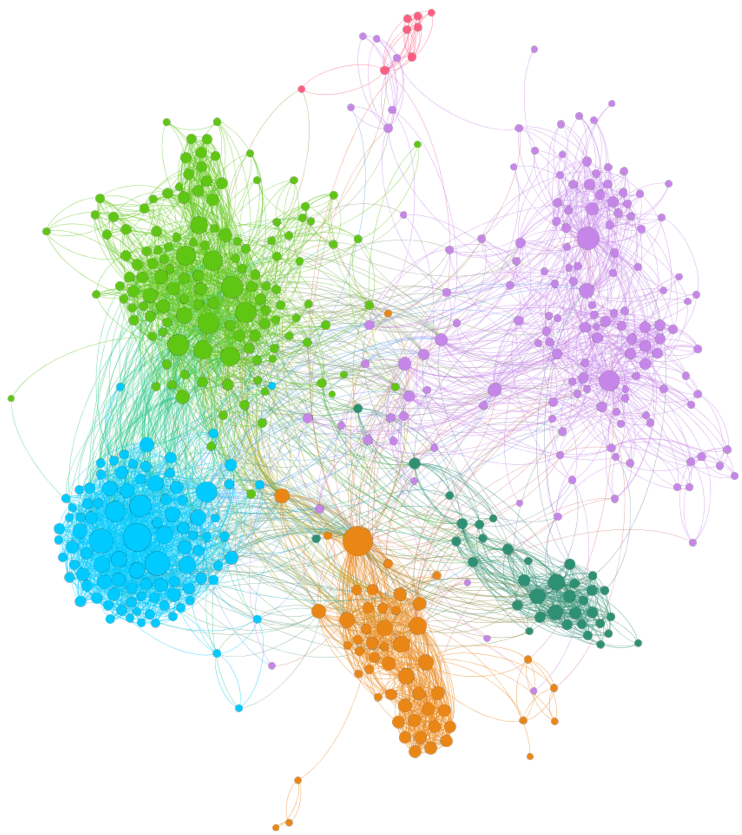




Purple color coding reflects the Hispanic Association of Colleges and Universities; light green coding reflects Joliet Patch; blue coding reflects Chicago Tribune; orange coding reflects Lewis University Recreation Fitness and Wellness; green coding reflects Lasallian link; pink coding reflects Chicago 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.

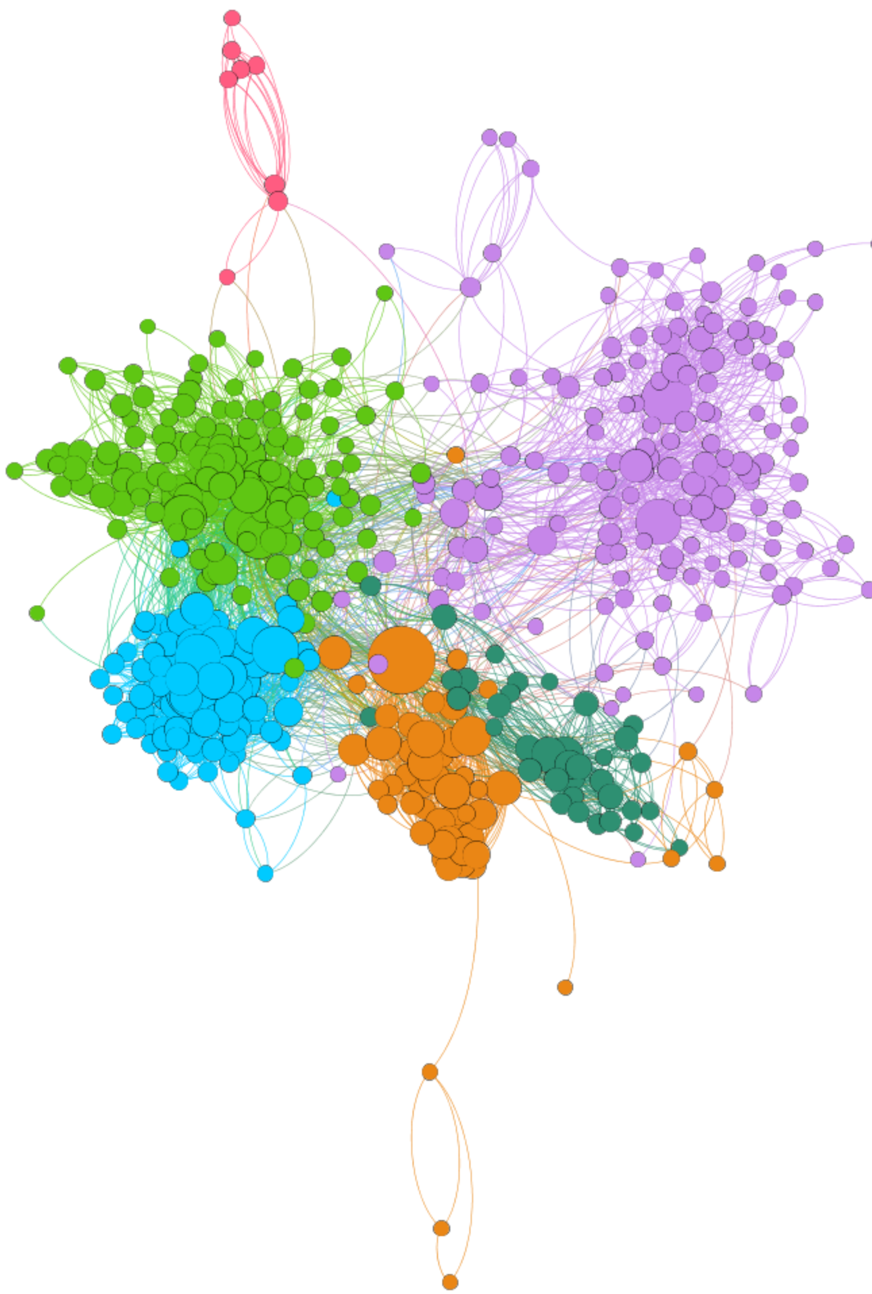
Also applied other layout algorithms to the graph such as ForceAtlas, ForceAtlas2 and Yifan Hu and the results are mainly the same. The only difference is on the dispersion of the graph as shown below. ForceAtlas, ForceAtlas2 and Yifan Hu gave more dispersed results compared to Fruchterman-Reingold layout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

ForceAtlas Layput Algorithm Results:



ForceAtlas Layout Algorithm Results:

Yifan Hu Layout Algorithm Results:

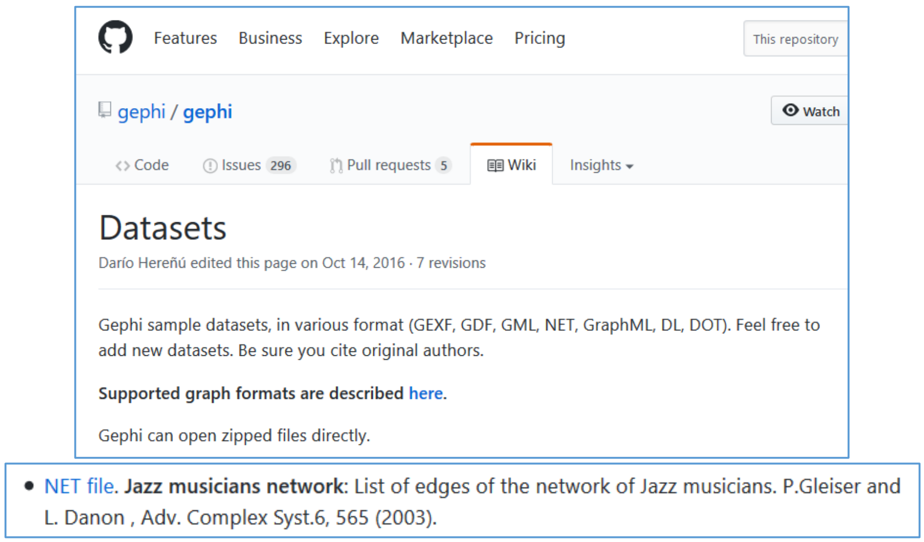


G. Conclusion

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

IV. 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.



A. Data Table

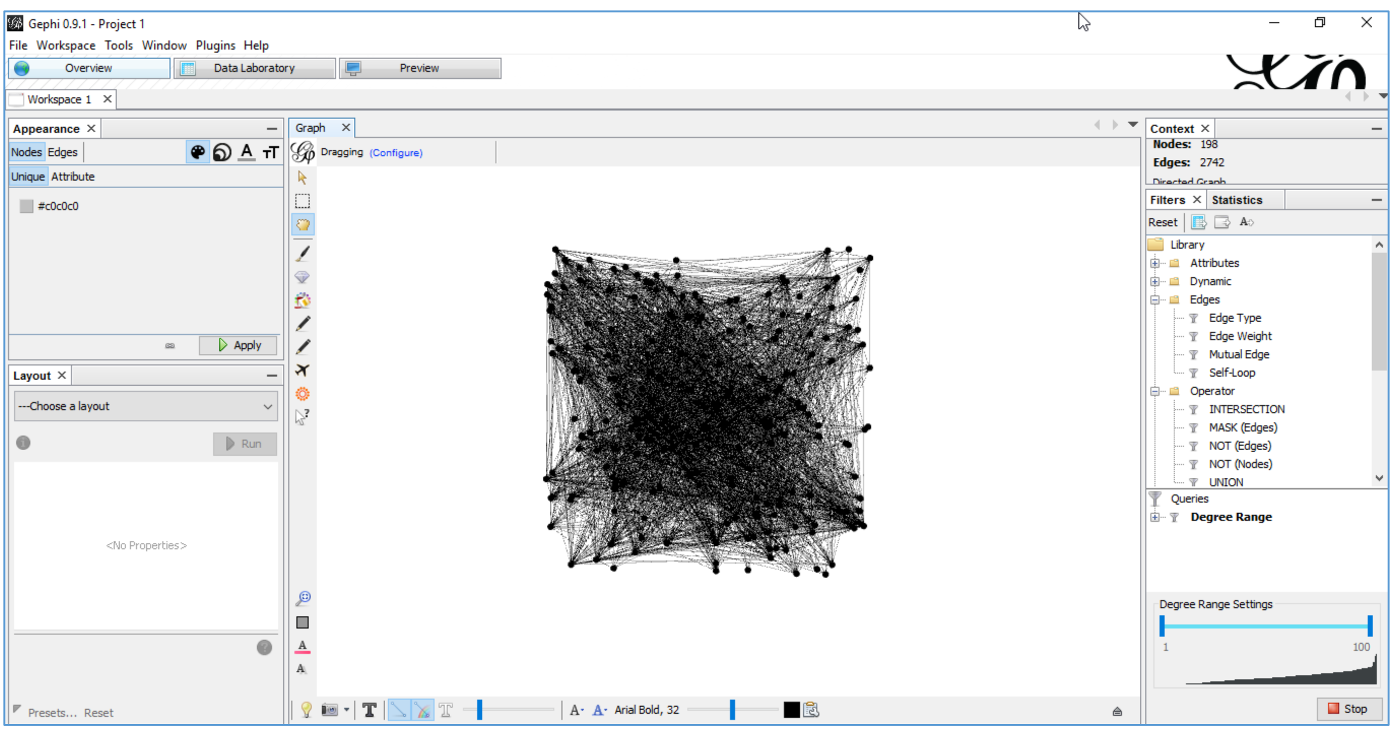
The data table provides the nodes and edges data. The nodes are unlabeled, identified only by an ID number (1 to 198). The edges have an ID, as well as a "source" and "target" ID for the directed edge (0 to 5433).

Data Table | Nodes | Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More actions

Source	Target	Type	Id
1	8	Directed	0
1	24	Directed	1
1	35	Directed	2
1	42	Directed	3
1	46	Directed	4
1	60	Directed	5
1	74	Directed	6
1	78	Directed	7
1	81	Directed	8
1	95	Directed	9
1	98	Directed	10
1	99	Directed	11
1	100	Directed	12
1	101	Directed	13
1	103	Directed	14
1	104	Directed	15
1	108	Directed	16
1	131	Directed	17
1	132	Directed	18
1	154	Directed	19
1	159	Directed	20
1	168	Directed	21
1	171	Directed	22
2	14	Directed	23

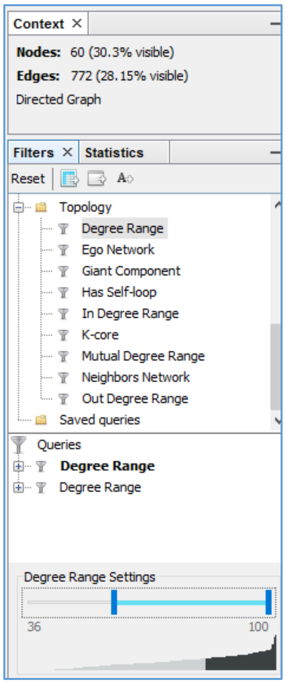
B. First Look at the Graph

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.



C. Filtering the Data

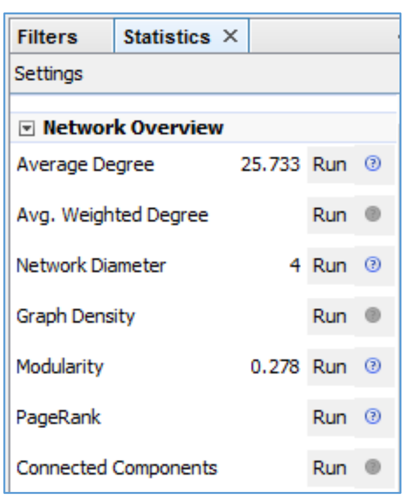
Filtering prunes the graph and keeps only the nodes that are relevant by setting various filtering parameters. Since this data table contained over 198 nodes and 2742 edges, used the degree range filter to remove nodes with degrees less than 36. Now the dataset is reduced to 50 nodes (30.3% visible)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



D. Getting Statistical Graph Data

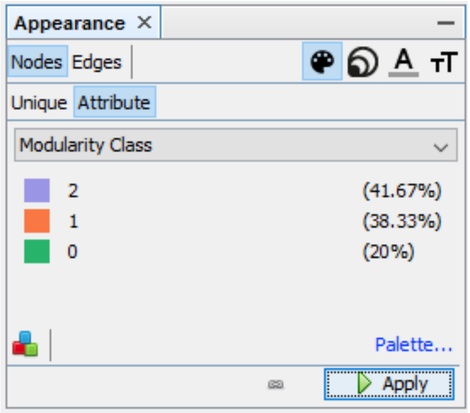
The following algorithms were run and values were computed.

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
    a. Betweenness centrality which measures how often a node appears on shortest paths between nodes in the network
    b. Closeness centrality which is the average distance from a given starting node to all other nodes in the network, and
    c. 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.



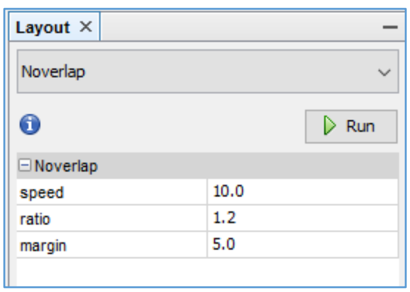
E. Appearance

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.

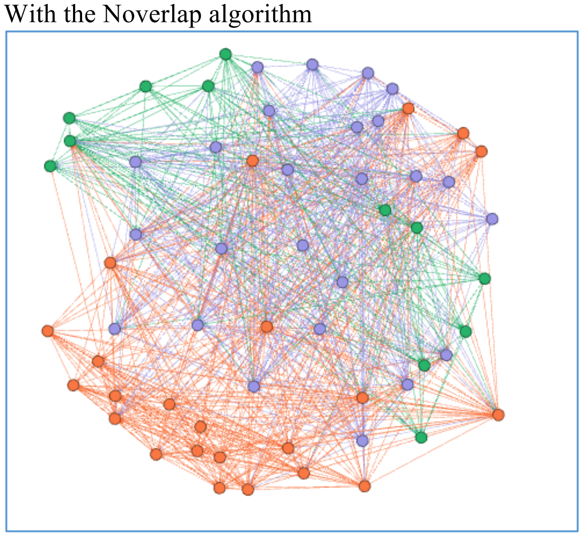


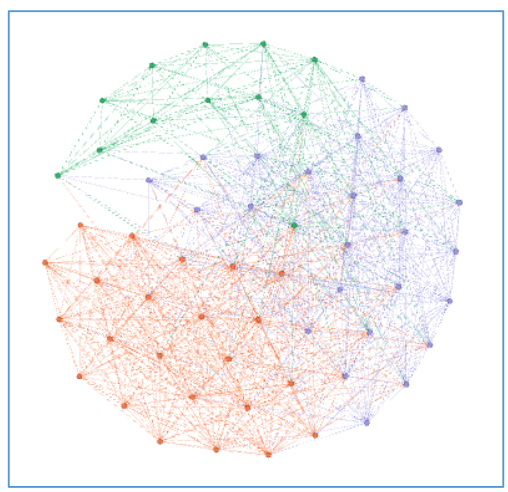
F. Layout Algorithms

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.



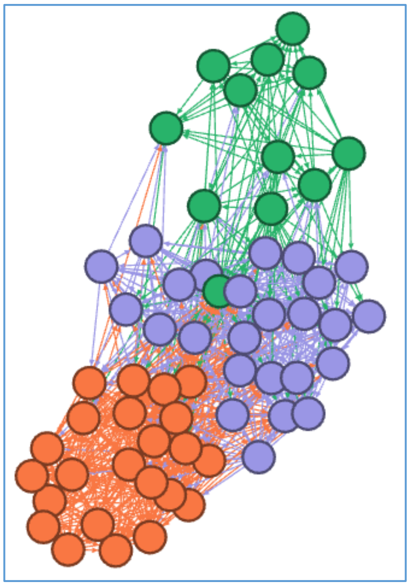
With the Noverlap algorithm





With the Yifan Hu layout algorithm.

With the Fruchterman Reingold 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