ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Social Network Analysis

Δανοπούλου Αιμιλία 3170033

Table of Contents

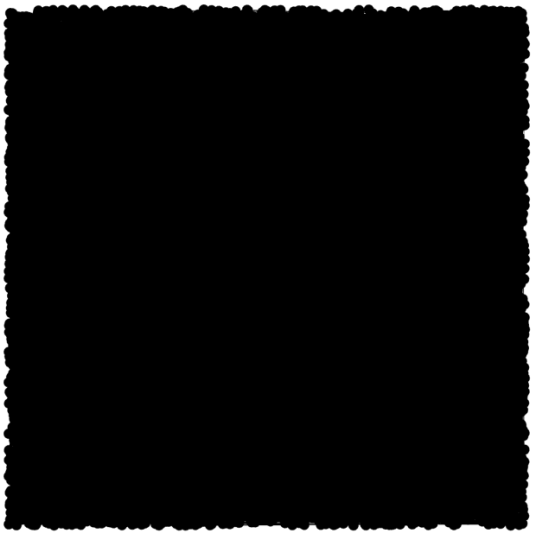
Overview

Astrophysics is a branch of science that focuses on applying the laws of physics and chemistry to explain the birth, life and death of stars, planets, galaxies, nebulas and other objects in our universe. This extraordinary science is studied by astrophysicists in their effort to give us a better understanding how the world outside our home Earth works. In the past decades there have been numerous publications regarding Astrophysics and in order to map their collaborations I decided to visualize the authors that have co-published scientific papers from 1995 to 1999.

The Dataset

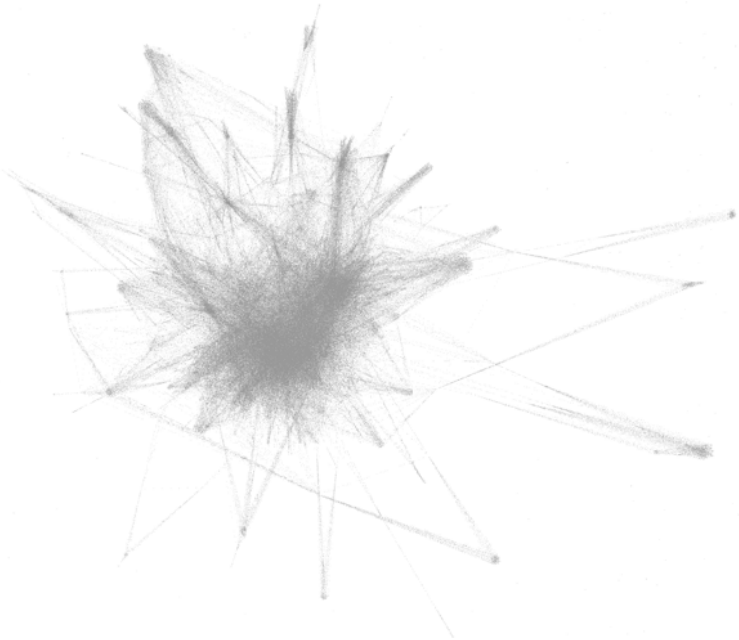
The Data set contains the collaboration network of scientists from 1995-1999
posting preprints on the astrophysics archive at www.arxiv.org, as compiled by M. Newman. It is an undirected dataset and contains weights. The network consists of nodes (vertices) V and edges E. The number of nodes in the graph is 16706 and the number of edges 121251.

I was lucky enough to find the dataset in a GML form. This means that the set did not need any kind of configurations and was ready for import.
After starting Gephi and importing the GML file this is how the dataset first appears, as a black square. It is obvious that our graph is far from its final form and needs configurations in order to look more like a real network. For this purpose, Gephi already provides a wide variety of algorithms and other means for visualizing the network.

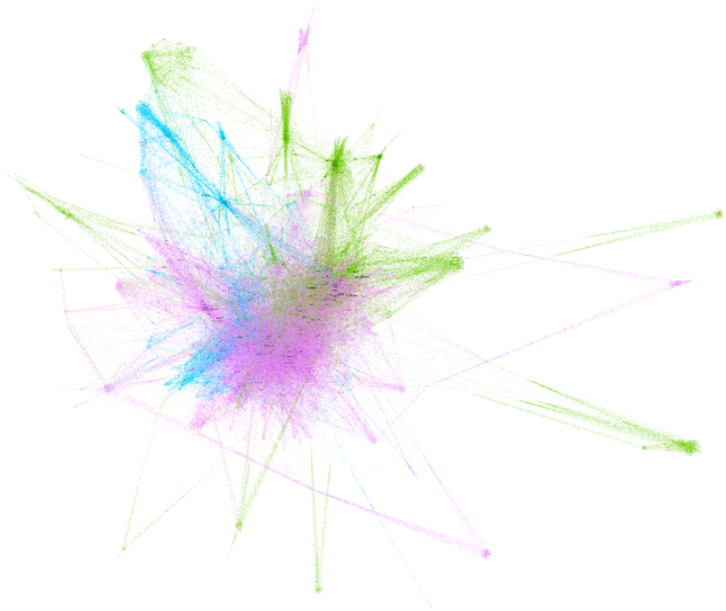


Main Visualization of Network

From the variety of algorithms that Gephi supports for layouts, we choose Force Atlas 2, a very powerful and precise algorithm.
After running Force Atlas 2 we now have a much clearer image. The graph is now expanded, and we can see some clusters of nodes that look like points. We proceed by calculating the average degree in order for the nodes size to appear according to its values. Finally, we run the Modularity algorithm in order to make some sense of the communities that exist. Given the number of nodes in the network, a resolution of 5 was adjusted for the Modularity in order to be able to differentiate between the communities. Using Modularity with a resolution of 1 would have generated a rainbow graph which would not give a lot of information regarding the communities.

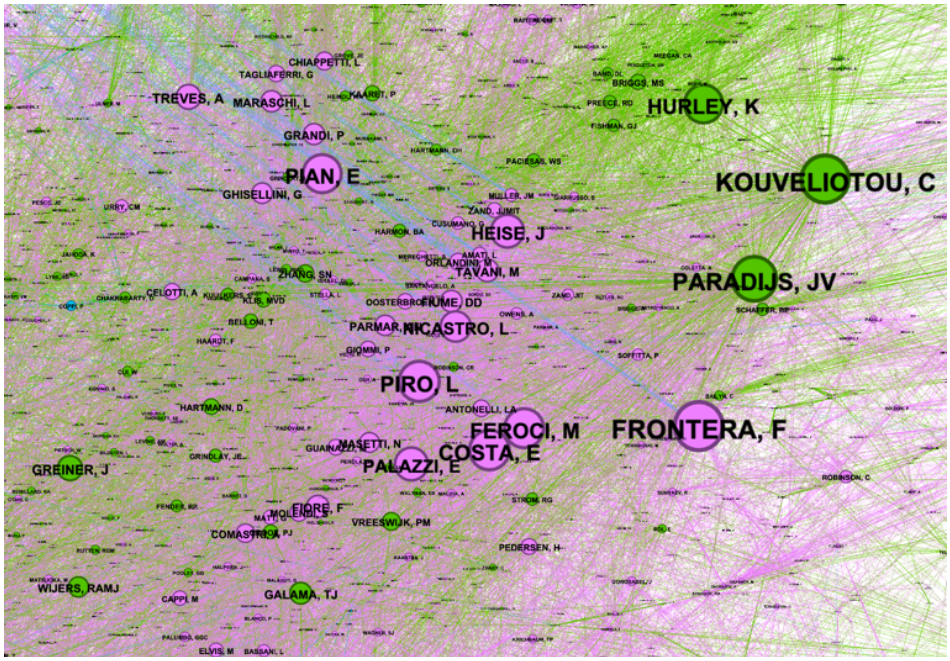


1st picture: Force Atlas 2



2nd picture: Degree ranking and Modularity

Zooming in the center of the graph we can clearly see the nodes with the highest degree with Frontera being the largest node, followed by Kouveliotou and Paradijs. Other noticeable nodes are Feroci M., Costa E., Piro L., Pian E. etc.



Very chaotic visualization of the center of the graph

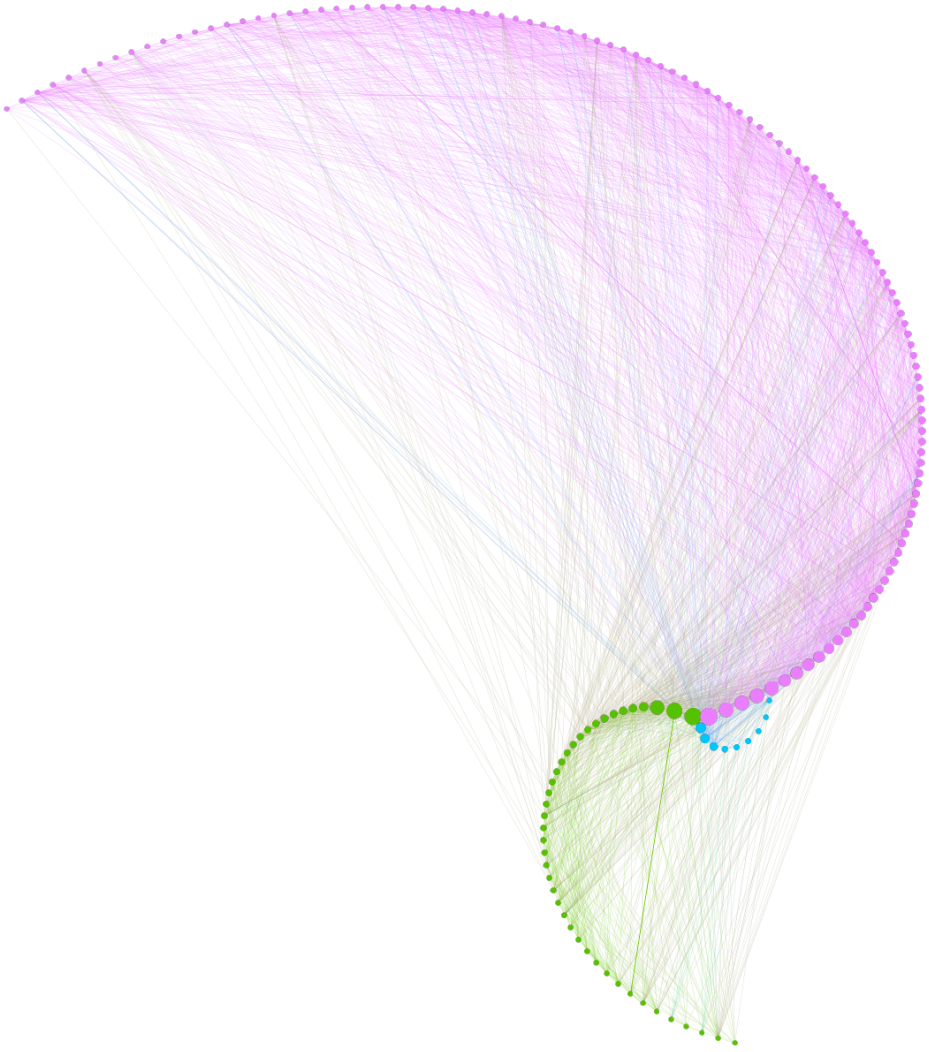
Since the original graph is still a mess with all these edges, I decided to apply some filtering for the metrics that follow.

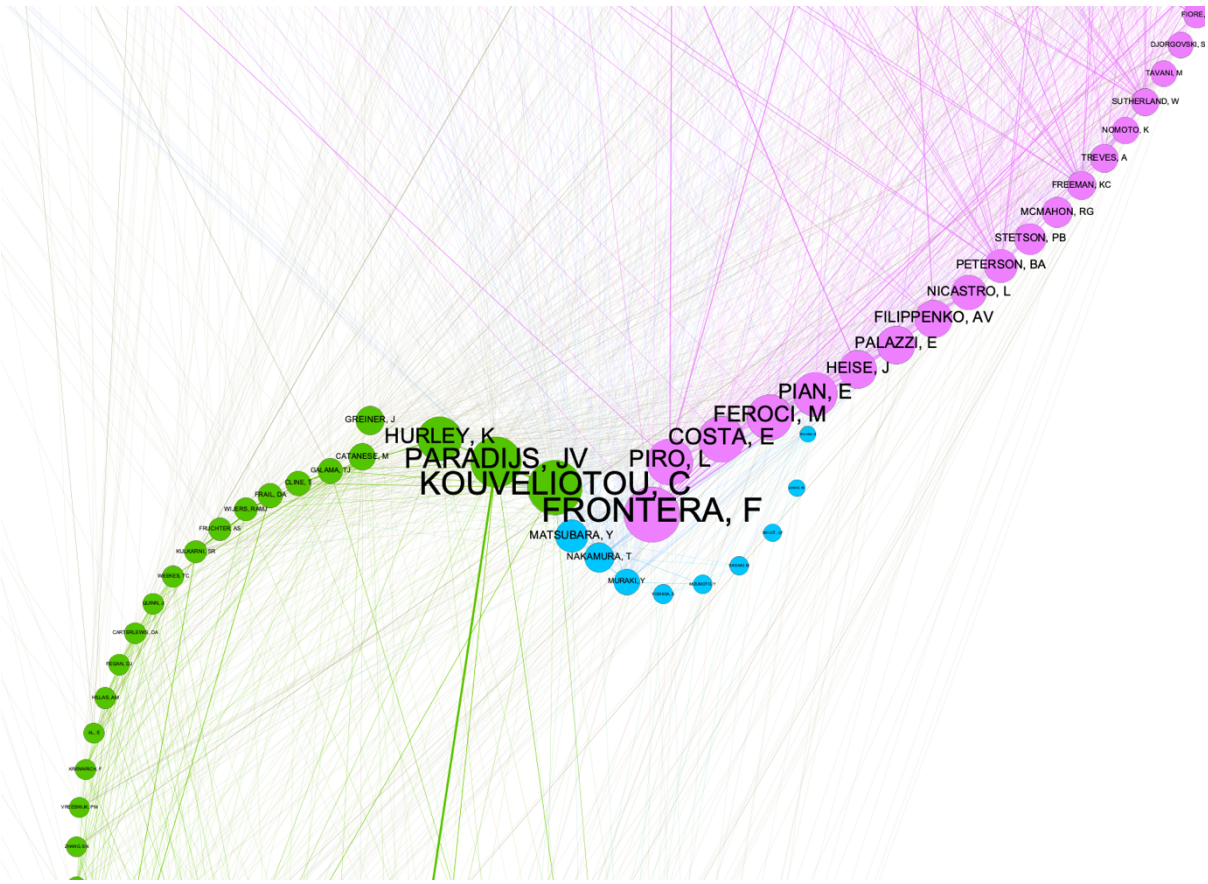
Modularity

Gephi provides an algorithm implementation of Modularity, which tries to provide some insight of the communities that exist in the graph. More specifically, it measures the strength of division of a network into modules.
Having applied the modularity algorithm and experimenting with it, I came to the conclusion that it was best to add a resolution of 5.0 to make the graph clear and comprehensive. In order to give further insight concerning the communities, I conducted a small research and came to some general conclusions. Specifically:
Blue Community consist mostly of Japanese astrophysicists and researchers.

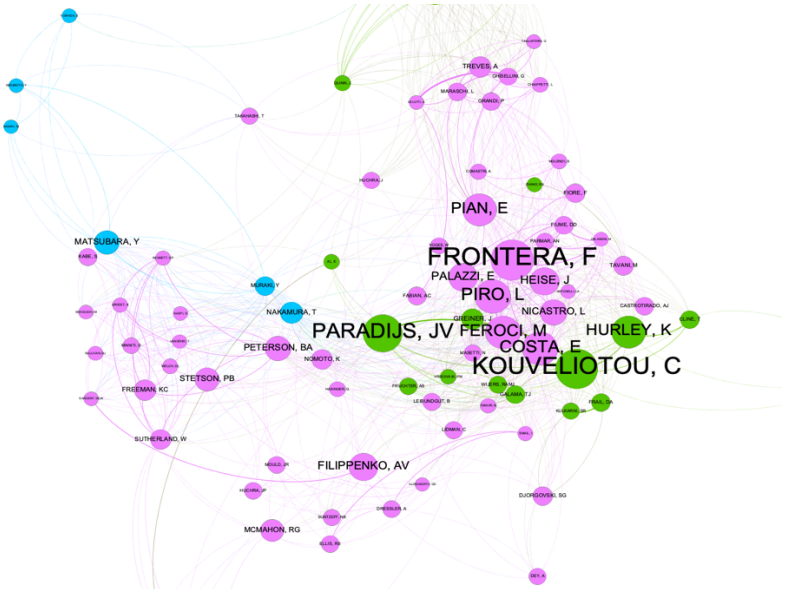
Green Community consists of people that attended or worked at Washington University of St. Louis and people that have connections from College of Dublin or have been there.
Purple Community consist mostly of Italians, English and Australian people, people part of Fermilab Cosmic Physics Center or people that took part in the Macho Project.
It is natural that in those 3 communities there have been mixed many different groups of people and consequently merged.

In the following picture the Radial Axis Layout was used with the nodes being grouped by Modularity Class and the order according to the degree measurements. This helps us see the differentiation between the 3 communities more clearly. Note that this picture does not contain all the nodes of the original network and have been filtered to display only nodes with a degree equal or higher to 100.





As seen from the above visualization many nodes from the purple community are connected to each other which is natural since the Modularity tries to color clusters of nodes that are more connected.



A graph of the Degree Ranking with Yifan Hu Proportional

Network Diameter and Average Path Length

The diameter of the network is 14. This means that the shortest distance between the two most distant nodes in our network is 14 hops.

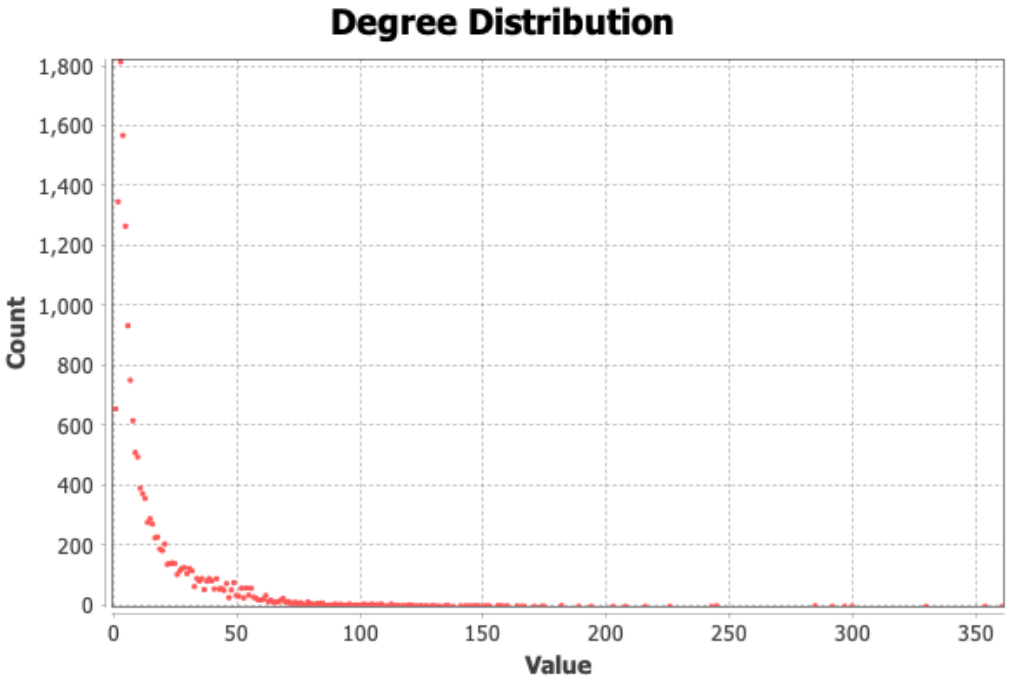
Results:

Diameter: 14
Radius: 0
Average Path length: 4.797959542913144

The average path length of the graph is 4.798 rounded. As it is clear, the average path length is significantly smaller in comparison to the diameter

Degree Measures

Average Degree: 14.516 Max Degree: 360 Min Degree: 1
Calculating the average degree, we can see each author's degree in the data laboratory. After sorting the degree column in descending order, we see that the node/author with the biggest degree is the Italian astrophysicist Filippo Frontera, followed by the Greek astrophysicist Chryssa Kouveliotou, followed by the Dutch astrophysicist Jan Van Paradijs. To my surprise Kouveliotou C. and Paradijs J.V. were married.

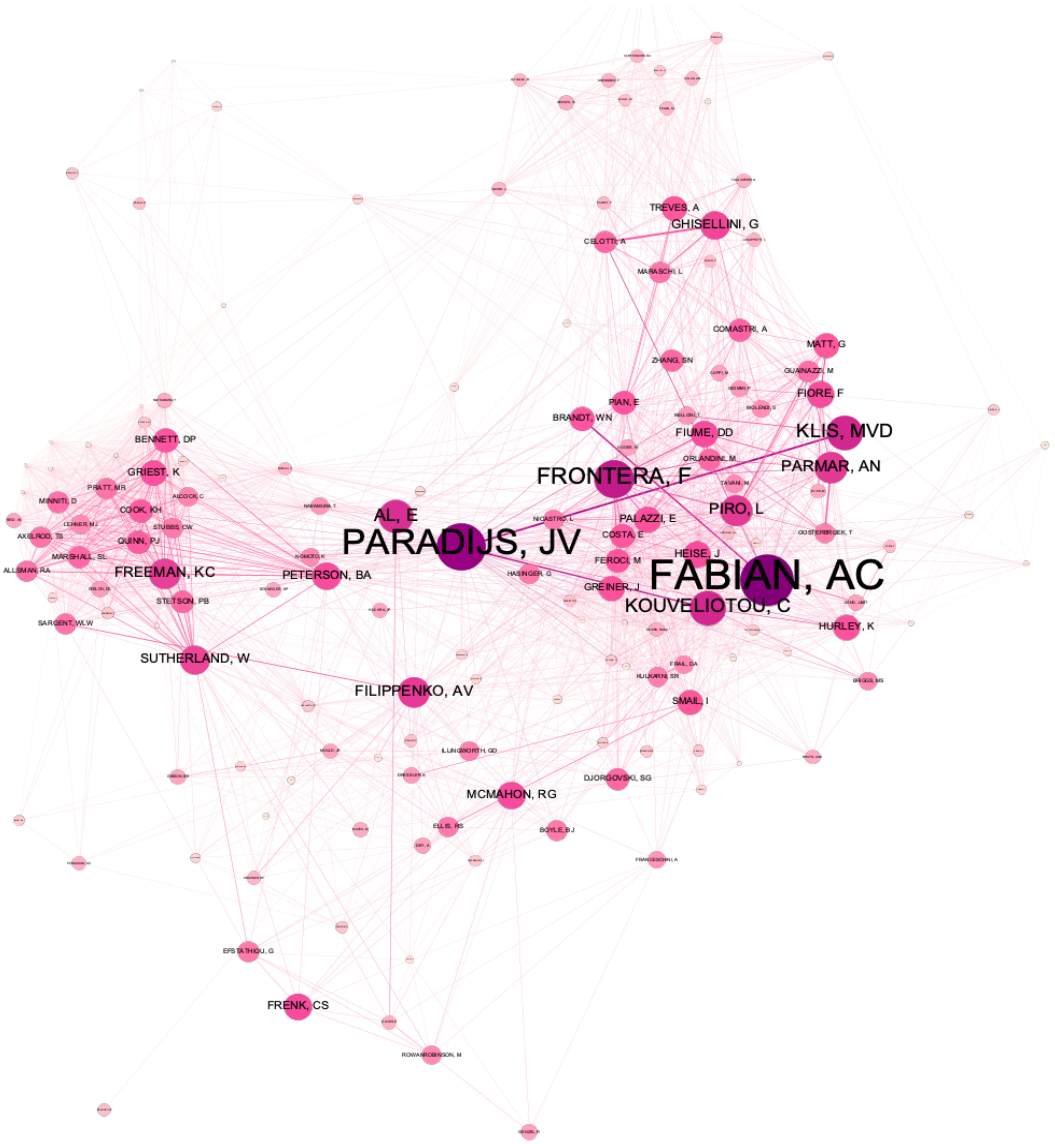


Id	Label	Interval	Degree
5502	FRONTERA, F		360
912	KOUVELIOTOU, C		353
1231	PARADIJS, JV		329
5507	PIRO, L		299
6197	COSTA, E		296
6199	FEROCI, M		291
1353	HURLEY, K		284
6216	PIAN, E		284
6198	HEISE, J		244
6201	PALAZZI, E		244
2338	FILIPPENKO, AV		242
6200	NICASTRO, L		225
217	PETERSON, BA		215
5782	MATSUBARA, Y		207
1796	STETSON, PB		202
468	MCMAHON, RG		193
435	NAKAMURA, T		188
210	FREEMAN, KC		181
230	GREINER, J		181
747	TREVES, A		181
2908	NOMOTO, K		174
222	SUTHERLAND, W		173
6755	CATANESE, M		170
231	TAVANI, M		166

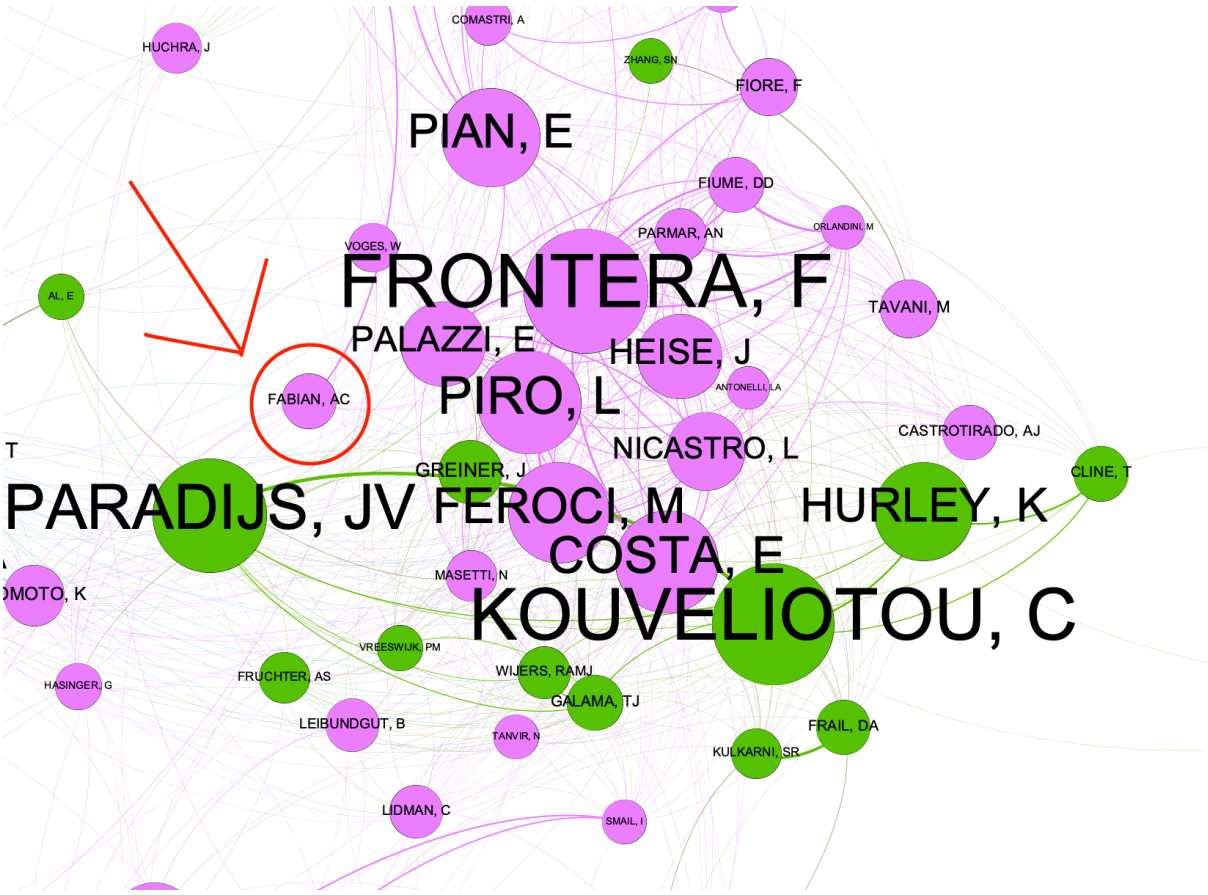
The degree of each node shows us the total number of relations they have. For example, Frontera has 360 relations/edges in total which means he has collaborated with 360 authors. Since our network is undirected, we don't have in-degree or out-degree measurements here.

The Dataset also provides weights and thus we can run the average weighted degree algorithm which returns as a result **4.099**.
Because the dataset doesn't mention what the weights represent, we can assume the weights on the edges as the number of times two authors have collaborated with each other, since in reality authors can collaborate numerous times.

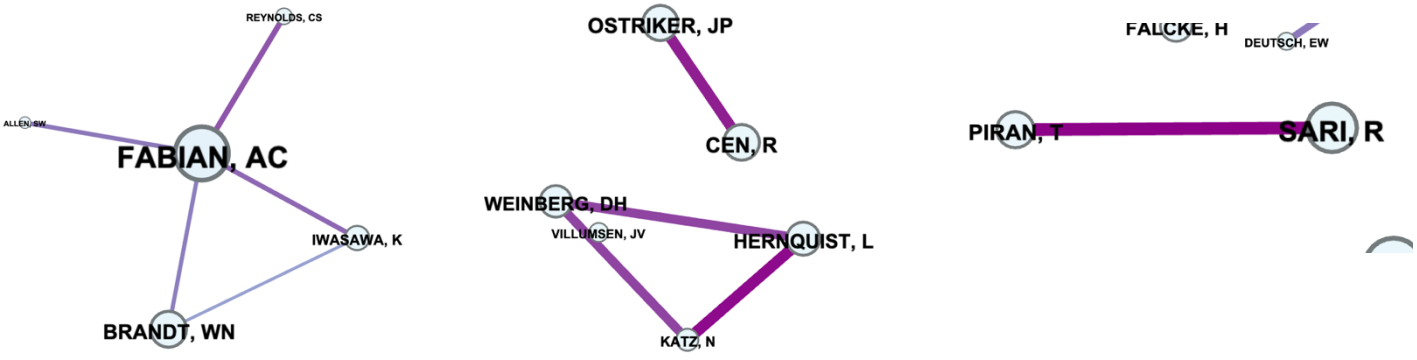
Weighted Degree Visualization follows with nodes ranking and color in accordance with the results.
Yifan Hu Proportional and Label Adjust were used for the layout as well as degree filtering from 100 to 360.



In the above picture the largest node is **Fabian AC.** having the biggest average weighted degree. This means that even though Fabian who did not appear very high in the Data laboratory of the degree measures (but still having a high degree of 156 and a decent node size in the graph), he appears to have collaborated many times with one or more of his neighbors.



Andrew Fabian as seen in the degree ranking visualization



After some filtering with **Edge weight range** and **Degree range**, we can see some of Fabian's most weighted edges, in order to understand how he came first in the previous ranking. Some other notable weighted edges appear in the next two photos.

Centrality Measures

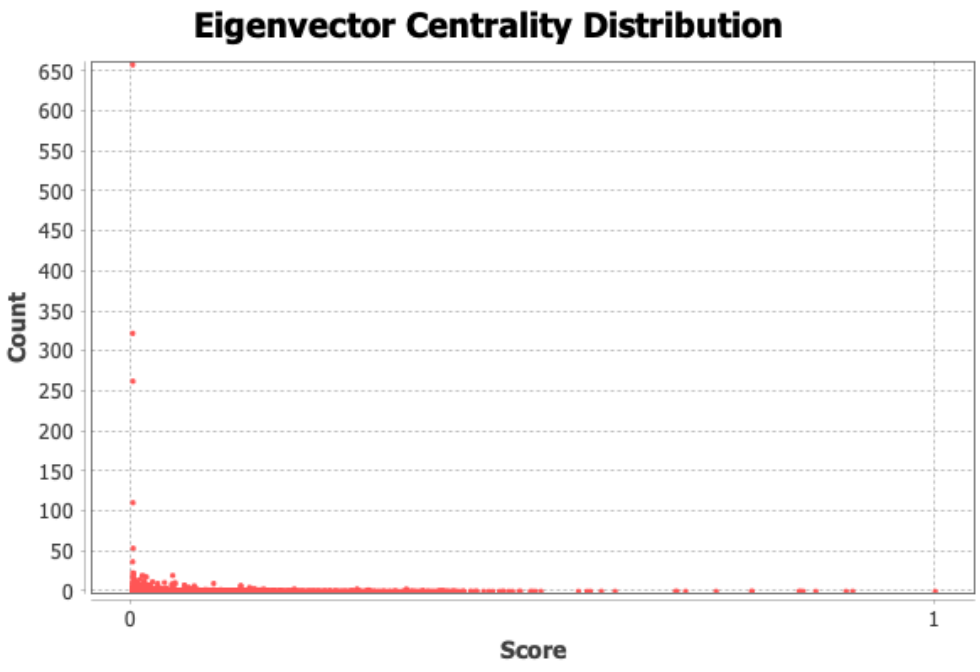
Eigenvector Centrality

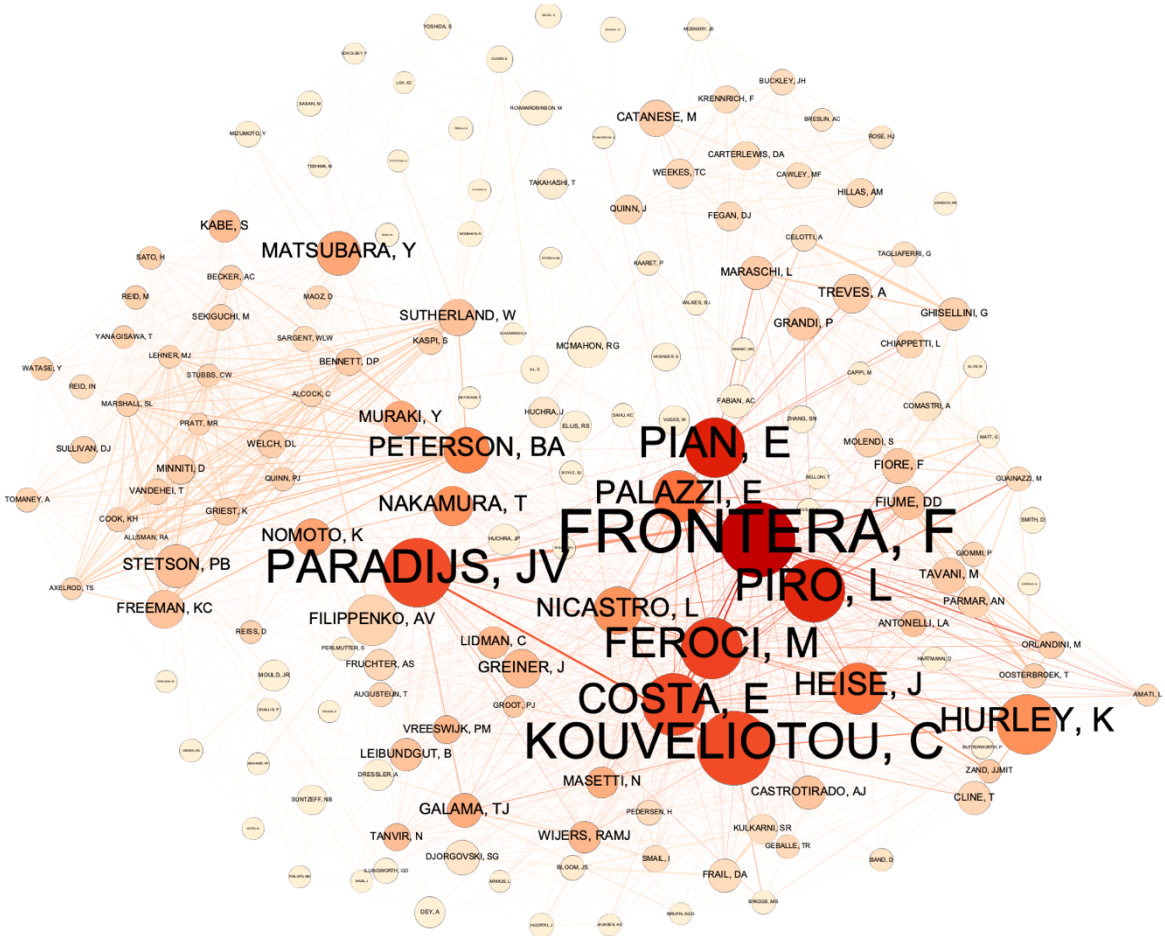
Eigenvector Centrality is a useful metric that measures the influence of every node in the network.

Label	... Eigenvector Centrality
FRONTERA, F	1.0
PIAN, E	0.897134
PIRO, L	0.888791
FEROCI, M	0.851179
KOUVELIOTOU, C	0.835268
PARADIJS, JV	0.832028
COSTA, E	0.830297
HEISE, J	0.771829
PALAZZI, E	0.771054
PETERSON, BA	0.726895
HURLEY, K	0.688907
NAKAMURA, T	0.678129
NICASTRO, L	0.676225
NOMOTO, K	0.601071
MATSUBARA, Y	0.583988
MURAKI, Y	0.570231
GALAMA, TJ	0.565996
MASETTI, N	0.555374
WIJERS, RAMJ	0.508577
VREESWIJK, PM	0.501653
GREINER, J	0.498256
LIDMAN, C	0.495439
KABE, S	0.483288
TANVIR, N	0.481383
FIUME, DD	0.473194
STETSON, PB	0.468465
SUTHERLAND, W	0.462236

In the Data Laboratory we can see in descending order the Eigenvector Centrality of the first 28 astrophysicists.

We can see that Filippo Frontera is the most influential node in our graph with a value of 1.0. Generally what Eigenvector Centrality tries to tell us is that nodes that are connected to an important node, are themselves also important, highlighting that what counts in not only the number of authors someone has collaborated with, but also with whom.





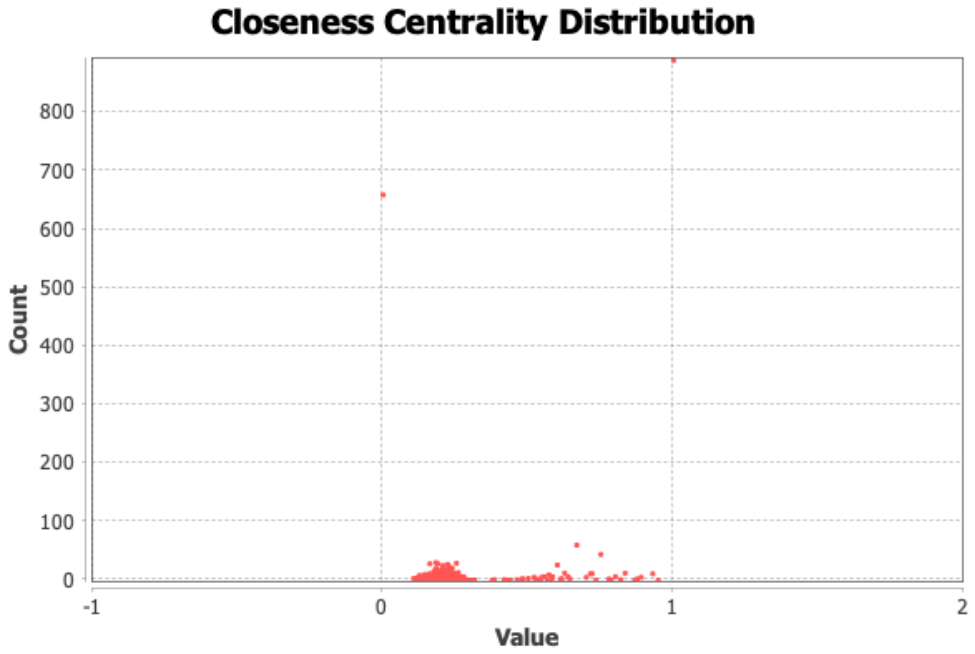
Eigenvector Centrality visualization using Fruchterman Reingold

In the above picture the coloring and nodes' size are according to Eigenvector Centrality. Note that the above picture depicts a subgraph with the nodes of ranging degree from 100 to 360.

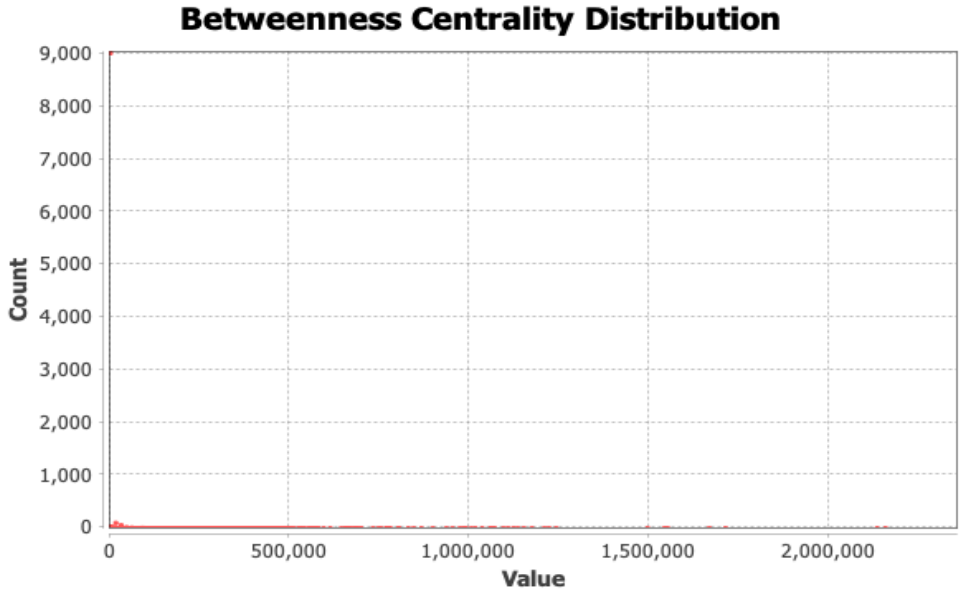
Closeness Centrality

Label	Closeness Centrality
FRONTERA, F	0.314765
KOUVELIOTOU, C	0.313946
PARADIJS, JV	0.312604
COSTA, E	0.311796
PIRO, L	0.311175
FEROCI, M	0.310908
PIAN, E	0.310187
FILIPPENKO, AV	0.308678
PALAZZI, E	0.307495
PETERSON, BA	0.303832
LIDMAN, C	0.303378
MCMAHON, RG	0.302939
GALAMA, TJ	0.302581
NOMOTO, K	0.300097
HEISE, J	0.300018
DJORGOVSKI, SG	0.299733
FRUCHTER, AS	0.299021
MASETTI, N	0.298889
GREINER, J	0.29812
FRAIL, DA	0.297989
WIJERS, RAMJ	0.297511

Closeness Centrality tries to measure how central a node is. The more central a node is, the closer it is to others. It is calculated as the average of the shortest path length from a node to every other node in the network. We should mention here that a node with a high degree tends to reduce one's average path and since closeness centrality is a measured based on it, it is only natural to conclude that closeness centrality and degree centrality are positively correlated. So, it is not a surprise that Frontera, Kouveliotou and Paradijs are in the top three, given their high degree.



Betweenness Centrality

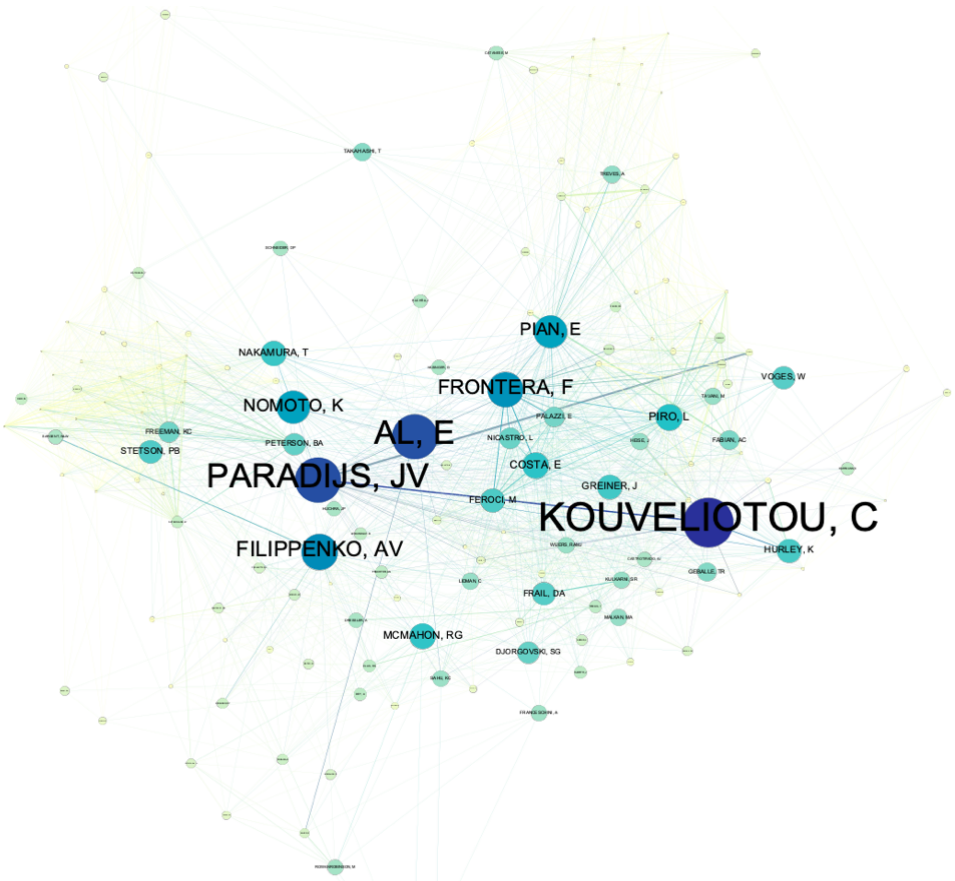


Label	Betweenness Centrality
KOUVELIOTOU, C	2358178.898407
PARADIJS, JV	2157279.359941
AL, E	2133589.971842
FILIPPENKO, AV	1711762.684111
FRONTERA, F	1666466.905479
NOMOTO, K	1551097.841968
PIAN, E	1541594.070621
BEAULIEU, JP	1494866.258015
PIRO, L	1240306.387861
COSTA, E	1220508.172269
MCMAHON, RG	1207355.631514
NAKAMURA, T	1172607.063435
GREINER, J	1150371.452647
SILK, J	1133393.703069
HURLEY, K	1121078.780457
FEROCI, M	1106145.084812
STETSON, PB	1093312.453233
FRAIL, DA	1068735.703206
VOGES, W	1065097.942458
KAMIONKOWSKI, M	1062072.403931
LIVIO, M	1058874.874886

Betweenness Centrality shows us which nodes, a pair of individuals need to pass from in order to reach each other.

In the left photo we see that Kouveliotou C. is the node from where every short path between a pair needs to traverse. We can assume then for example, that if an author from the edge of the network wants to collaborate with an author from the other edge, it would be easier to first reach Kouveliotou and through her connections to reach the target.

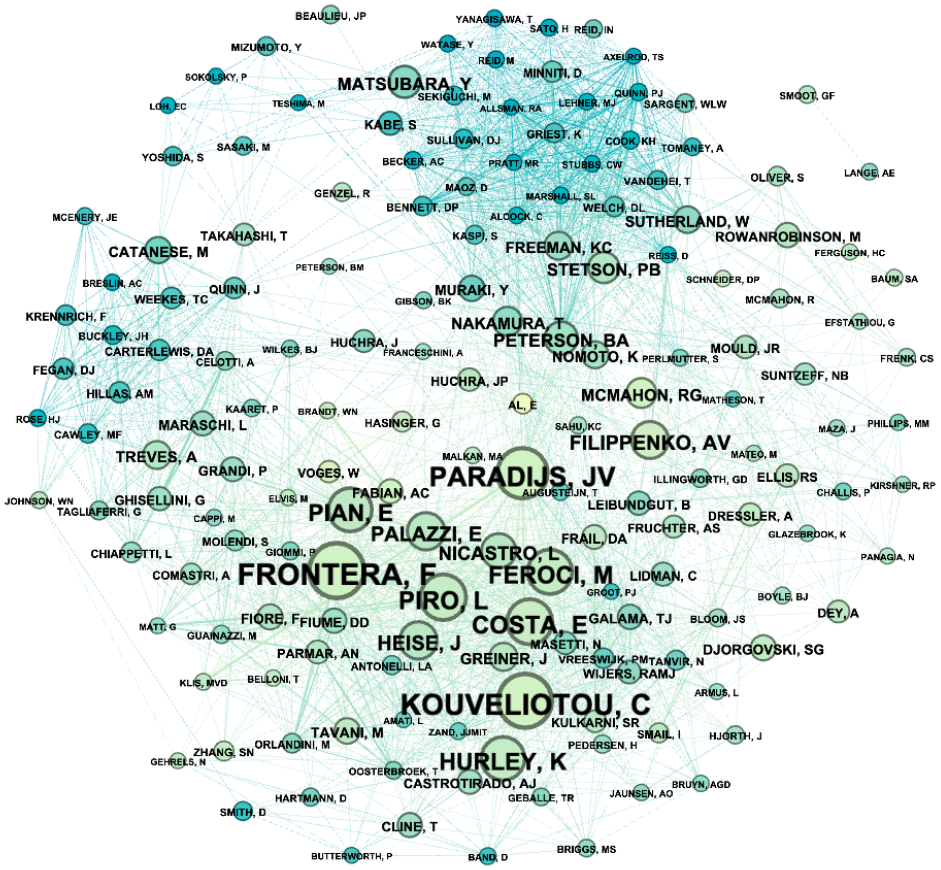
Removing nodes with high betweenness could cut the graph into multiple unconnected components.



Visualization with Betweenness measures coloring and node ranking

Clustering Metrics

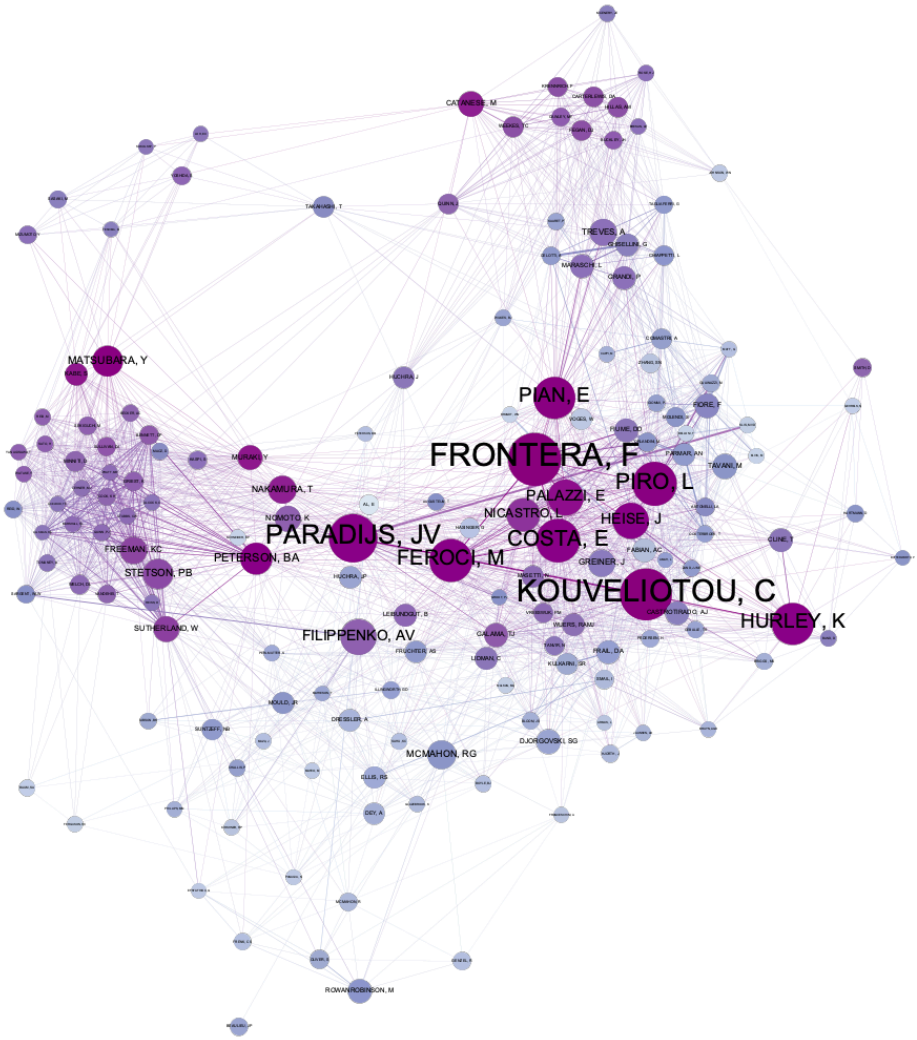
The clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. In order to examine this measure in our dataset we run the Average Clustering Coefficient algorithm for all the nodes, which returns the result of **0.726.**
In the visualization that follows we see a portion of the original graph in order to make a better understanding of the clustering effect. The darker green nodes here have a higher coefficient and are mostly connected with other darker nodes meaning that they tend to cluster together creating a very clean interconnected “mass” of nodes. On the other hand lighter colored nodes such as Frontera with a low clustering coefficient measure don’t have the tendency to cluster with other nodes and thus don’t belong to any kind of “mass”.



Fruchterman Reingold with Clustering Coefficient rank coloring and Degree rank

Number of Triangles

The total number of triangles in the network is **756019.** Generally, a triangle means that there are 3 nodes that each connect only with two of the others, creating a cycle. In the following graph the nodes have been ranked and colored according to their number of triangles. In order for Gephi to calculate the number of triangles in a graph, it uses clustering coefficient and thus the results are given together when running the Clustering Coefficient algorithm.

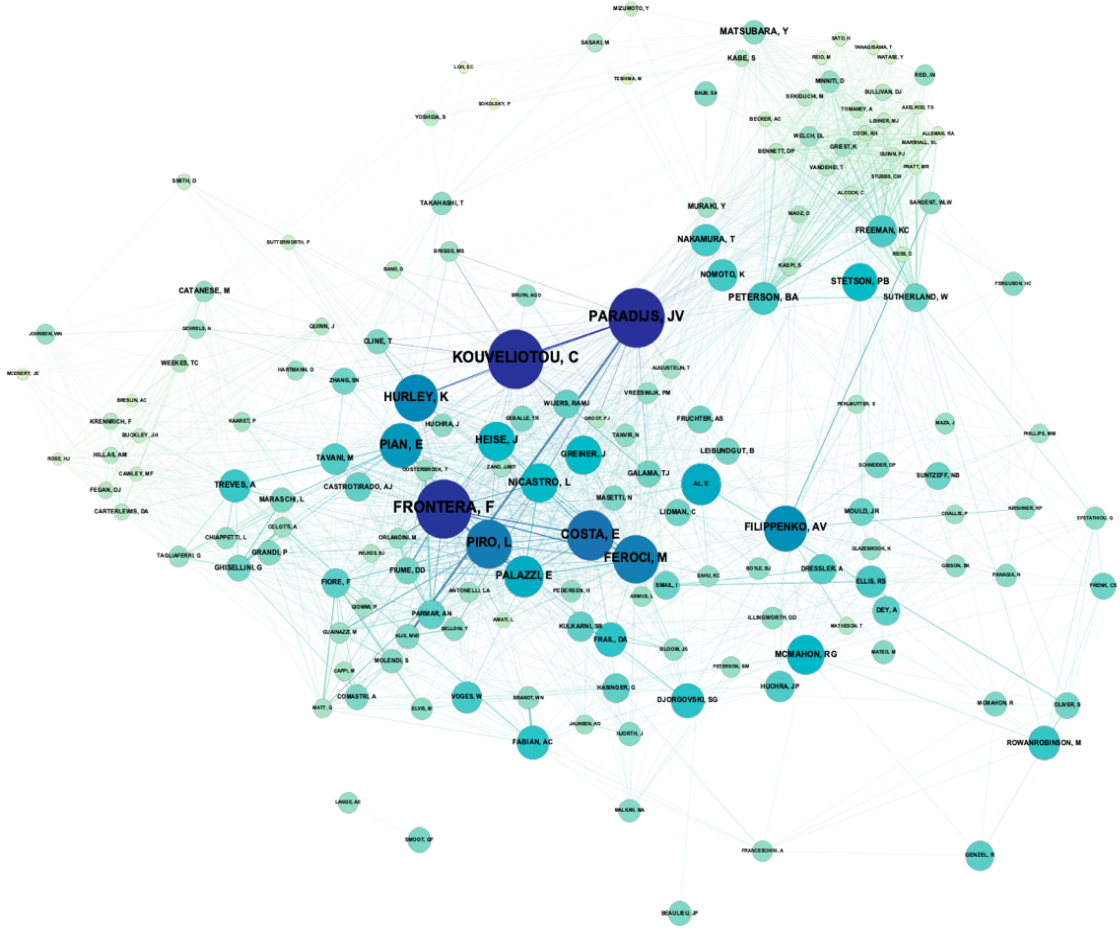


Label	Number of triangles
KOUVELIOTOU, C	6266
MATSUBARA, Y	5597
PIAN, E	5575
PIRO, L	5407
FEROCI, M	5178
COSTA, E	5072
PARADIJS, JV	4977
HURLEY, K	4954
PETERSON, BA	4719
HEISE, J	4662
PALAZZI, E	4389
KABE, S	4280
MURAKI, Y	4268
CATANESE, M	4216
NAKAMURA, T	4183
NICASTRO, L	3834
SUTHERLAND, W	3592
FREEMAN, KC	3468
STETSON, PB	3415
CARTERLEWIS, DA	3352
HILLAS, AM	3352
FEGAN, DJ	3348
WEEKES, TC	3329
KRENNRICH, F	3258
SULLIVAN, DJ	3213
BENNETT, DP	3109

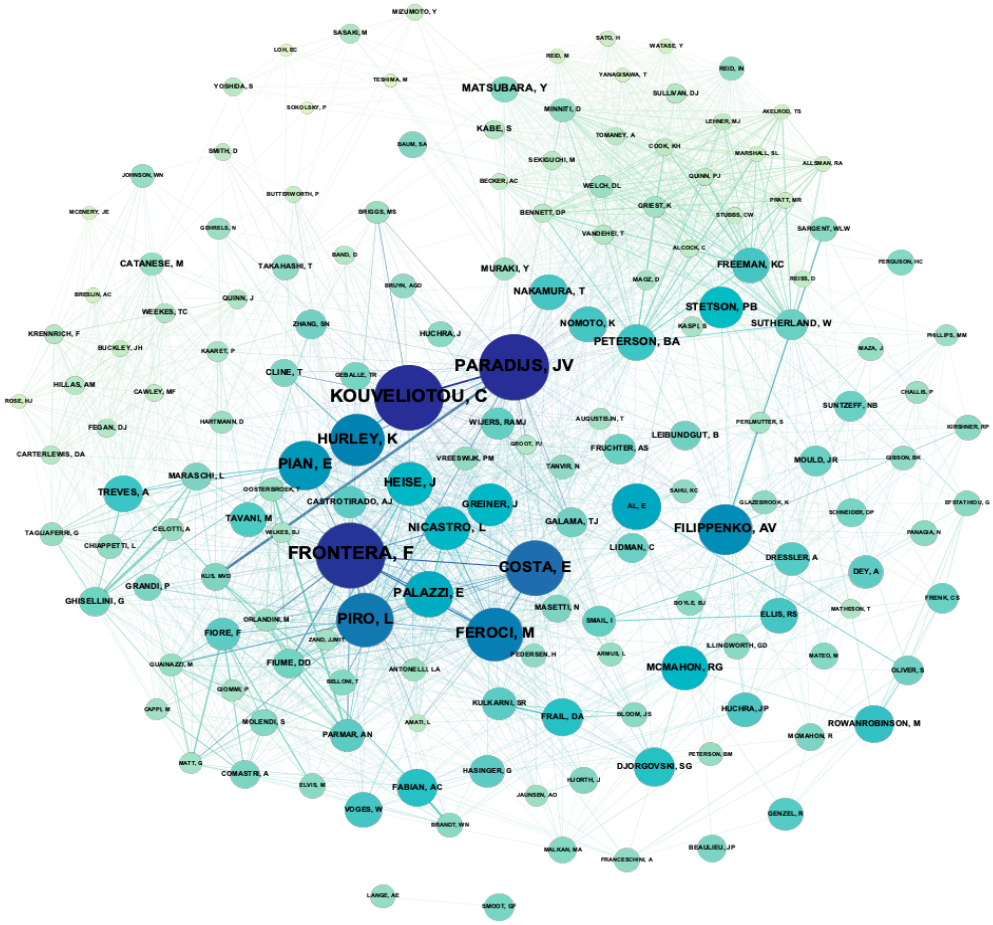
The global Cluster coefficient is an effective way to calculate how many triangles exist in our network since it is based on triplets of nodes.
From the data laboratory we see that Kouveliotou has the highest number of triangles, which means that there are many nodes, for example A and B, that she connects to and that an A-B edge also exists. In real life, this tells us that authors that have collaborated with Kouveliotu have also collaborated with each other.

PageRank Algorithm

The main question in this case was if the PageRank algorithm could be applied for such a dataset and how it would perform. Even though PageRank was originally created by Google in order to measure the importance of websites in a directed network, it can also be used in undirected networks. The algorithm was performed for the whole graph but here we see a part of it.
Configurations for the visualization were:
1.Range 100 – 360 degree and only 1% of nodes visible
2.Yifan Hu Proportional with 70 strength
3.The coloring and node's ranking for this subgraph was based on the PageRank metrics.



1st visualization of PageRank



The same subgraph with Fruchterman Reingold and adjusted gravity to 6

Label	...	PageRank
PARADIJS, JV		0.000803
KOUVELIOTOU, C		0.000797
FRONTERA, F		0.000794
COSTA, E		0.000672
PIRO, L		0.000656
FEROCI, M		0.000648
HURLEY, K		0.000631
FILIPPENKO, AV		0.000615
PIAN, E		0.000597
AL, E		0.000561
PALAZZI, E		0.000556
MCMAHON, RG		0.000533
HEISE, J		0.000527
NICASTRO, L		0.000524
GREINER, J		0.000519
STETSON, PB		0.000503
DJORGOVSKI, SG		0.000467
FRAIL, DA		0.000458
FABIAN, AC		0.000458
ROWANROBINSON, M		0.000452
TREVES, A		0.000445
PETERSON, BA		0.000436
NOMOTO, K		0.000431
NAKAMURA, T		0.00043

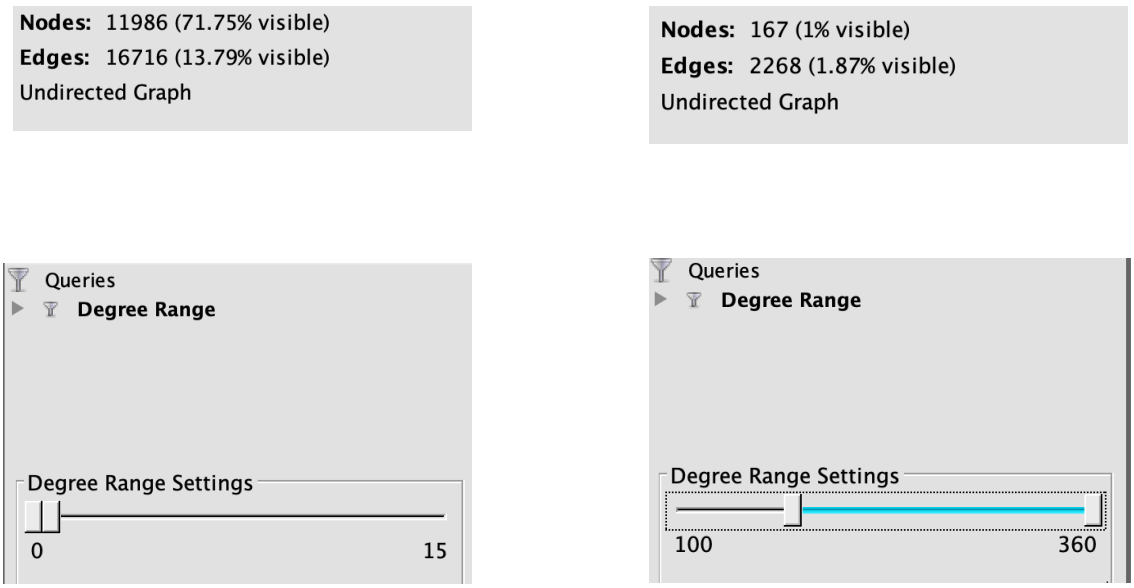
In the data laboratory we see in descending order the most important nodes. PageRank being a variation of Eigenvector Centrality still keeps Frontera in 3rd place in comparison to 1st place, as we see in the data laboratory of the Eigenvector results. A very small and maybe insignificant difference in positions that still shows us how important Filippo Frontera is in our network. The same applies to Kouveliotou and Paradijs that generally placed high as well in the previous metric.

Density

The Density of the network is 0.001 or $8.68953161589233x10^{-4}$ as provided by Gephi and NetworkX respectively.
Density is calculated by dividing the total edges that already exist in the network m with the total possible edges that could be created.

$$UndirectedNetworkDensity = \frac{TotalEdges}{TotalPossibleEdges} = \frac{Cardinality}{Size} = \frac{m}{n(n-1)/2}$$

A **low-density** network generally means that there are many authors in the graph that haven't collaborated with many others. This is expected, since filtering the graph for a degree range of 0 – 15, a total 71.75% of nodes are visible and 13.79% of the edges. This implies that the majority of nodes have a very small degree which means that they don't have many edges/connections to others. This is another way for confirming that the graph is of low-density.

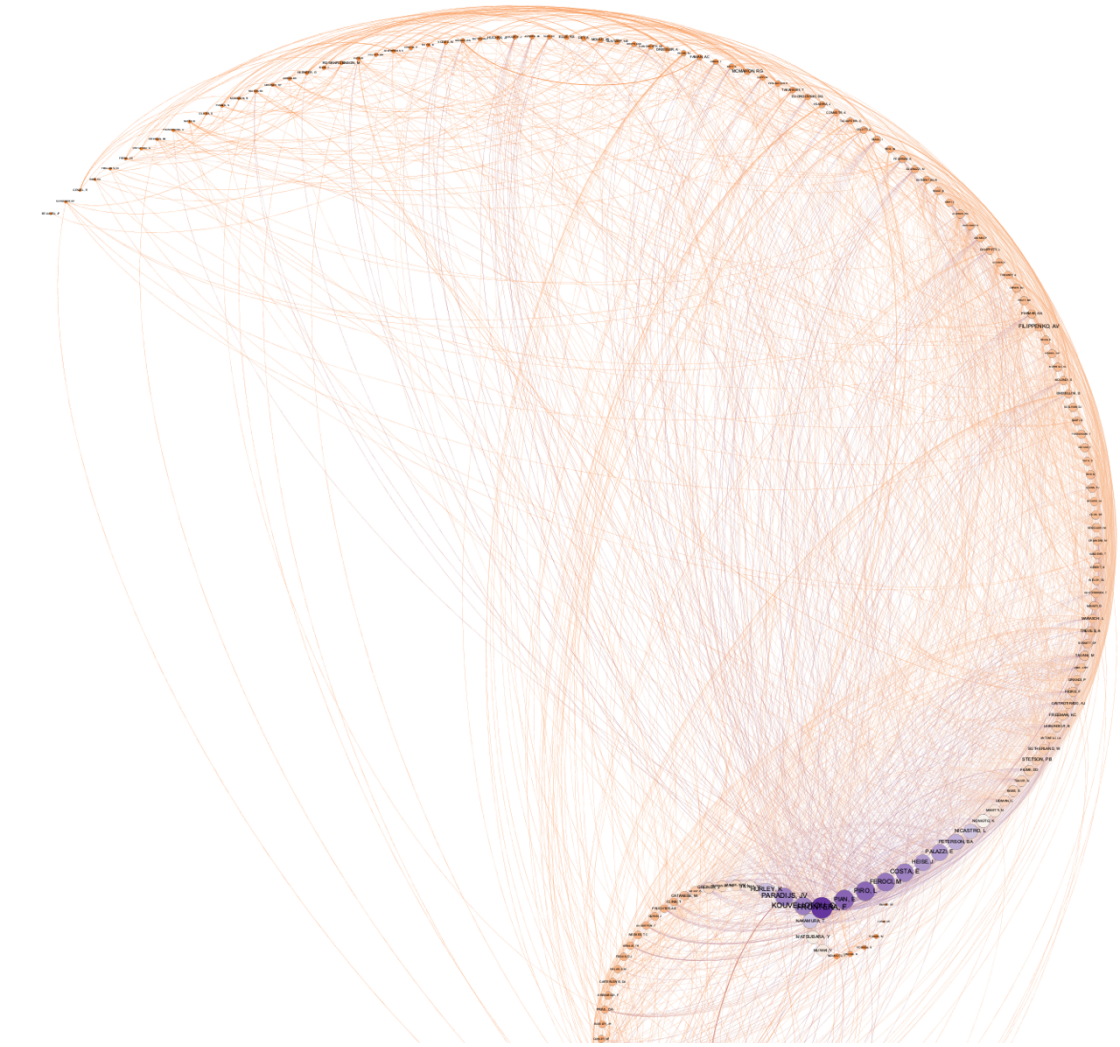


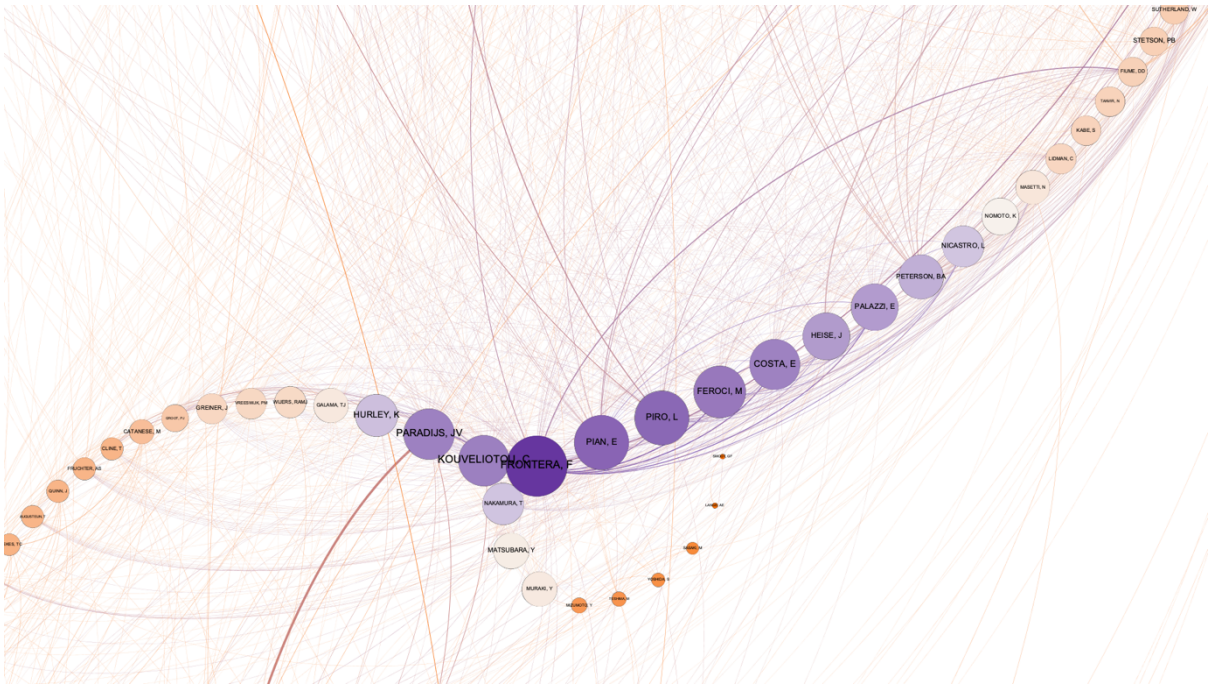
Comparison of visibility between degree range 0 to 5 and 100 to 360

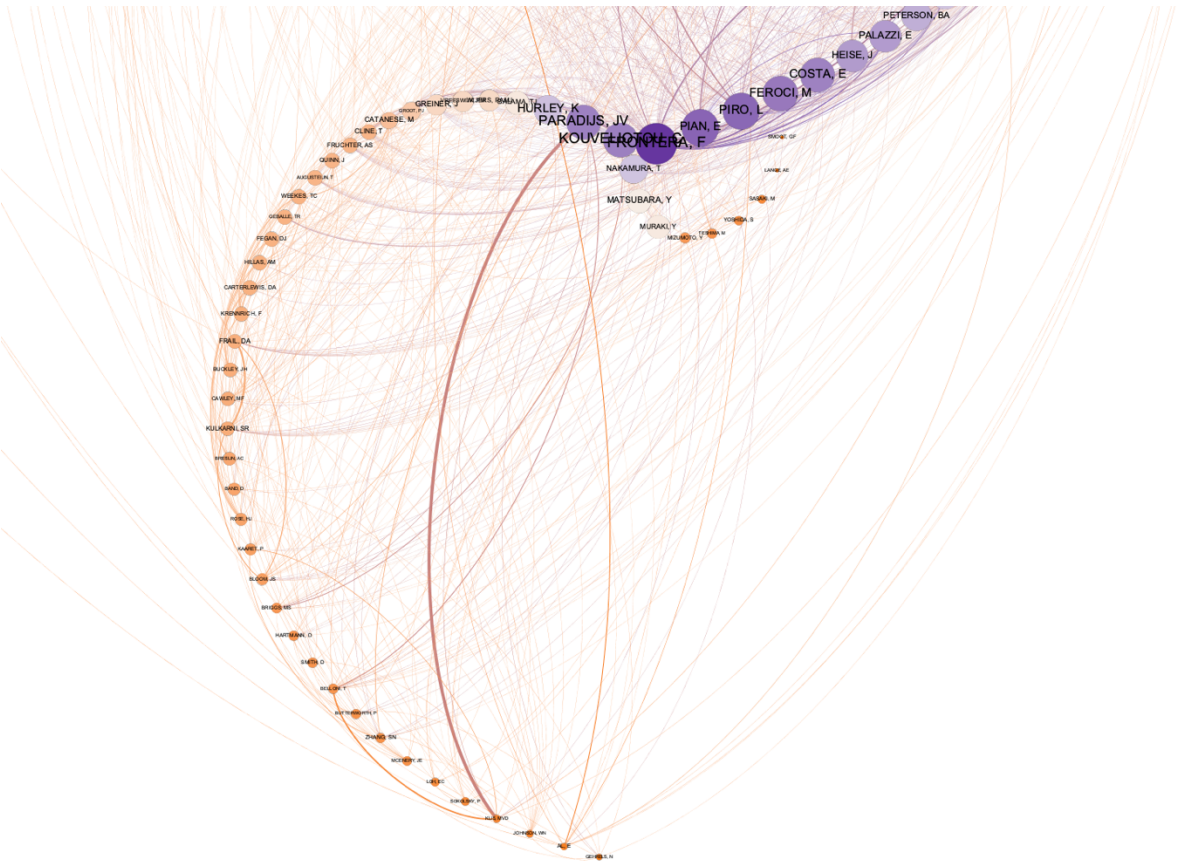
Homophily

Homophily is way of examining the relationship between nodes given a specific characteristic. For this case I examined the relation between nodes given their Eigenvector Centrality measures. For the visualization, the Radial Axis Layout was used once again where the nodes were grouped by Modularity, while the order of nodes in the sparse, the coloring and ranking were according to the Eigenvector Centrality.

While looking at the pictures below, we can see that the purple nodes (the one's with the highest eigenvector value) have mostly edges between other purple nodes, while orange nodes have mostly edges with other orange ones. This was to be expected since we have explained previously that nodes with a high eigenvector value are connected with other important nodes (thus other purple).

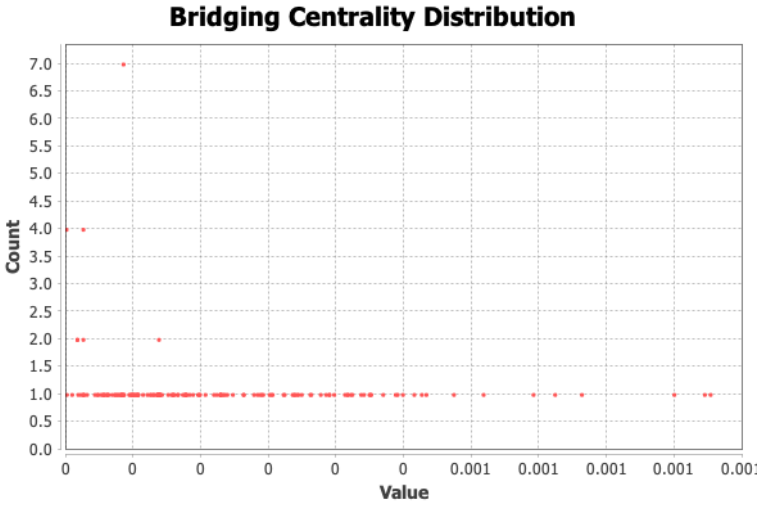


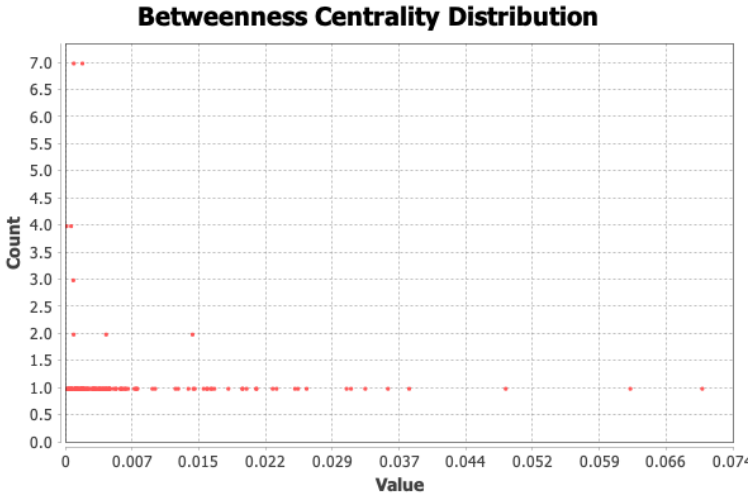

SUTHERLAND, W
STETSON, PB
FIUME, DD
TAMARI, N
KABE, S
LIDMAR, C
NAGETTI, N
NOMOTO, K
NICASTRO, L
PETERSON, BA
PALAZZI, E
HEISE, J
COSTA, E
FEROCI, M
PIRO, L
PIAN, E
FRONTERA, F
KOUVELIOTOU, C
PARADIJS, JV
HURLEY, K
GALAMA, TJ
WIJERS, RAMJ
GREINER, J
CATANESE, M
CLINE, T
NAKAMURA, T
MATSUBARA, Y
MURAKI, Y

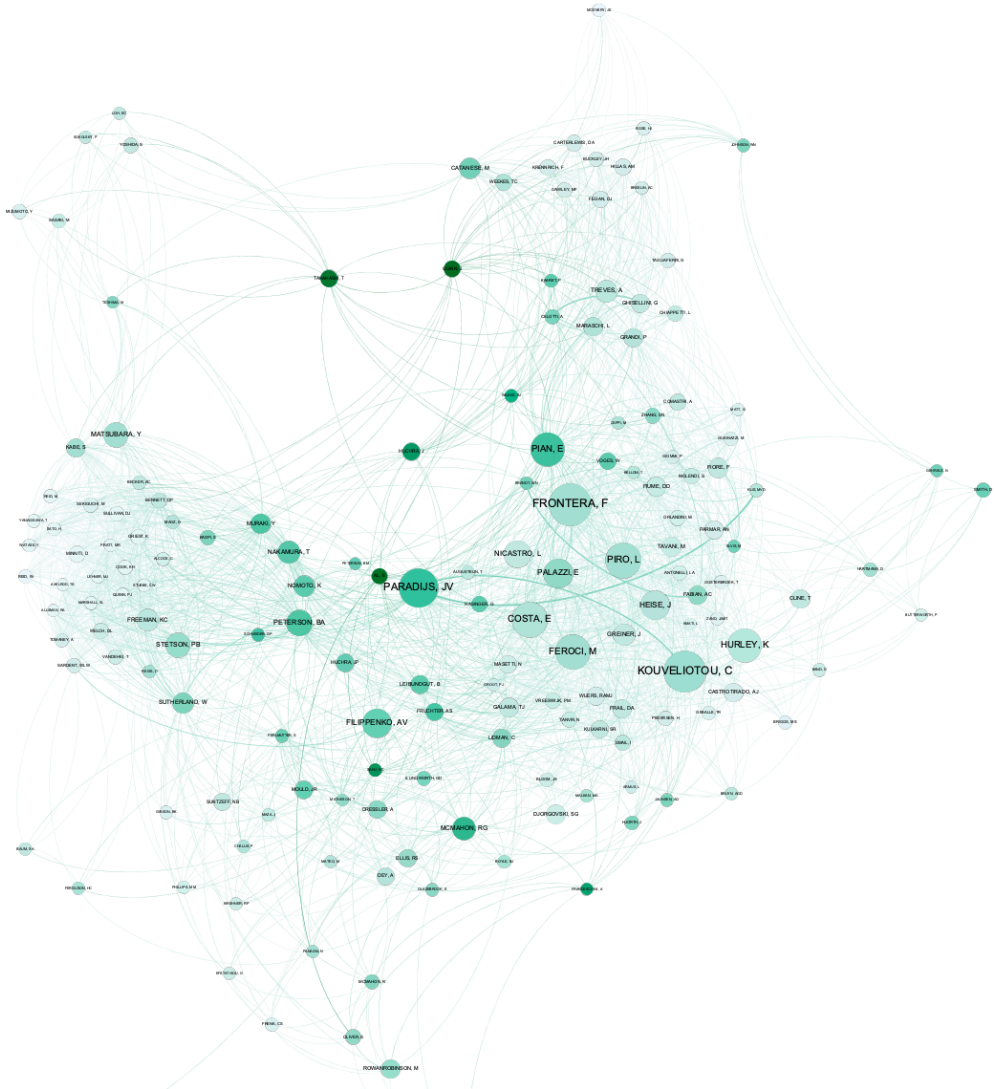


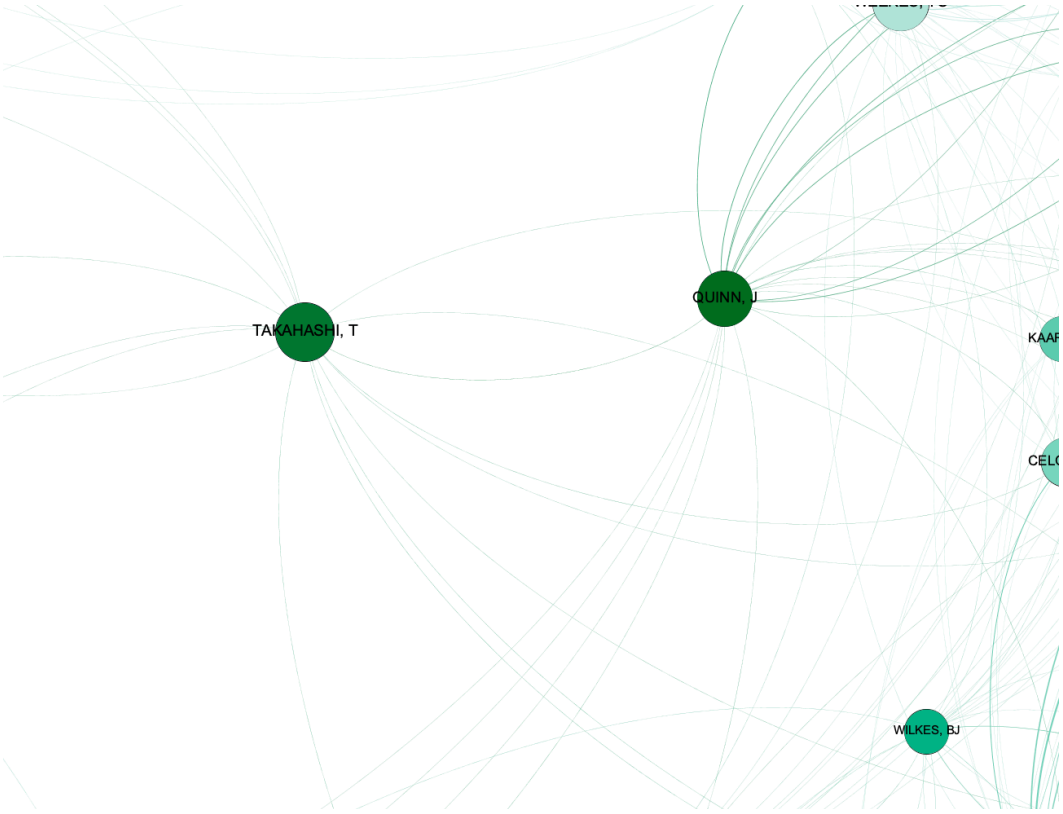
Bridges and Local Bridges

Bridges in graph theory are essentially edges whose deletion can increase the graph's number of connected components. In order to study the bridges that exist in the network
Gephi provided a plug-in called Bridging Centrality. After downloading it and running it **only** for nodes with a degree from 100 to 360 we have the following Bridging Centrality Distribution and Betweenness Centrality Distribution. It should be **noted** here that the Bridging Centrality algorithm wasn't executable for the whole graph since Gephi would freeze. For this reason, I decided to analyze Bridging Centrality only for nodes with a range of degree 100 to 360.


Bridging Centrality Distribution
Count
Value
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
0
0
0
0
0
0.001
0.001
0.001
0.001
0.001


Betweenness Centrality Distribution
Count
Value
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
0.007
0.015
0.022
0.029
0.037
0.044
0.052
0.059
0.066


PIAN, E
FRONTERA, F
PARADIJS, JV
COSTA, E
PALAZZI, E
PIRO, L
NICASTRO, L
FEROCI, M
KOUVELIOTOU, C
HURLEY, K
PETERSON, BA
FREEMAN, KC
STETSON, PB
SUTHERLAND, W
FILIPPENKO, AV
MCNAMARA, RG
MATSUBARA, Y
NAKAMURA, T
HEISE, J
GREINER, J
TREVES, A



From the data laboratory and as seen in the above visualization which was made in accordance with the Bridging Centrality results, Quinn J. and Takahashi T. are the nodes with the highest values. Here these two nodes have many edges that are local bridges, which means that most of their endpoints have no friends in common. Deleting their edges would then increase the distance between any pair of nodes that previously connected through one of them. Here we can also examine the triadic closure, which is a tendency of nodes that share a common neighbor to connect themselves. For example, here if node A and node B are connected to Quinn, then there is a high probability that in the future an edge from A-B would form.