Katholieke
Universiteit
Leuven

**Department of
Computer Science**

COVID-19: CO-AUTHOR NETWORK ANALYSIS

Siddharth Agarwal (r0773458)
Evangelia (Elina) Oikonomou (r0737756)

Academic year 2019–2020

Contents

1 Introduction

SARS-CoV2 took the world by storm late last year and has since been having the outsized effect on all social, economic, and cultural activities that we have come to recognise historically of a pandemic. This has also led to a sharp increase in the amount of research and development in the domain of understanding the virus, its effects on people, the ways in which it spreads, and the ways we can slow its spread, cure Covid-19, and develop a vaccine for it.

With the veritable explosion of Covid-19 research content, arose the problem of classifying it, retrieving relevant information form the literature, and discovering how the research is being conducted across groups and domains. One of the best known ways to gauge how researchers are collaborating is a co-authorship network.

Co-authorship networks are defined as follows: [1]
Let U be a set of bibliographic units (journal and conference publications, books, and so on), and let A be the set of authors appearing in U. The co-authorship network corresponding to U is most commonly defined as an undirected and weighted graph $G = (V, E)$ with the following properties:

- The set of nodes V corresponds to the set of authors A, i.e. each author from A is represented by one node in G.
- Co-authorship networks are author-centered onemode projections of bipartite networks linking researchers to bibliographic units they (co-)authored.

Note also that a co-authorship network is not the same as a citation network.

2 Dataset

We are using the CORD (Covid-19 Open Research Dataset) provided by the Allen Institute for Artificial Intelligence. [3] The dataset is updated frequently, and we stuck to using the version released on April 10_{th}. This version has 5162 articles authored by 20531 researchers.

We use the metadata.csv file from the dataset as our data. This consists of sixteen columns that list out for each paper, titles, reference numbers, list of authors, pubmed ids, time of publishing, journal, etc. Of these we selected the paper reference number and the author names for the graph.

While exploring the dataset, we came across some quality issues that we have listed below:

- The dataset contains not just academic research but also news articles from scientific reporting websites.
- The listed authors for some articles in the dataset are not always the people who have written the articles. Some articles just list the journal or website where they were published.

2.1 Exploratory Data Analysis

Before we began the graph analysis, we analysed the data for trends, patterns, and outliers. We found trends for number of authors per paper, the number of papers per author, the most productive authors, and papers that were a result of relatively large collaborations.

On calculating the authors per paper, we see the following trend curve:

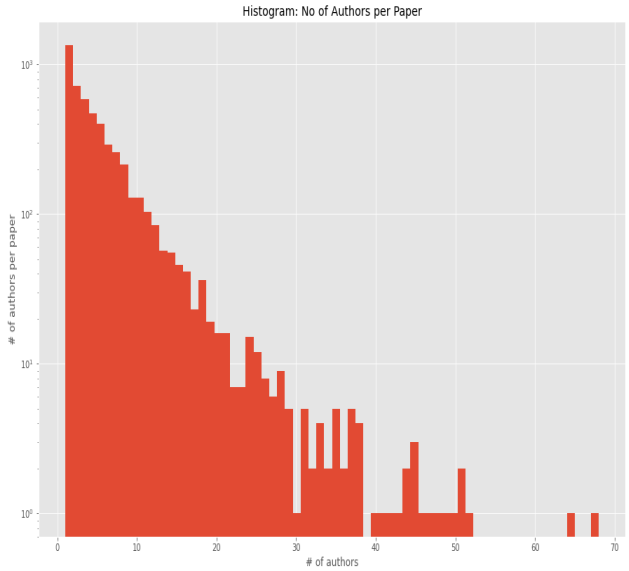


Figure 1: Number of authors per paper

We obtained the following statistics for the dataset:

- Maximum authors per paper: **68**
- Min authors per paper: **1**
- Average number of authors per paper: **5.21**
- Standard deviation: **5.85**

Number of Authors per paper	Number of Papers	%age of papers
1	1343	26.02
2	722	13.99
3	590	11.43
4	474	9.18
5	399	7.73
> 5	1634	31.65

We can see that a very large percentage of papers have been authored either by a single author or by groups of more than five authors.

On calculating the papers per author, we see the following trend curve:
Similarly we calculate statistics for author productivity:

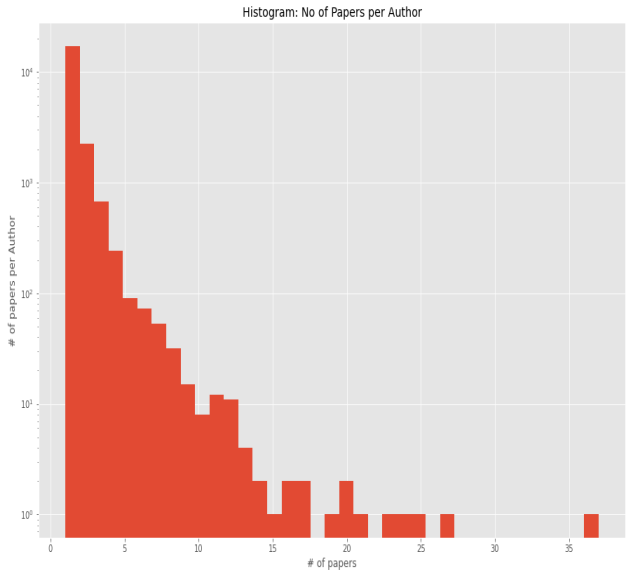


Figure 2: Number of papers per author

- Maximum authors per paper: **37**
- Min authors per paper: **1**
- Average number of authors per paper: **1.31**
- Standard deviation: **1.06**

Number of papers per Author	Number of Authors	%age of Authors
1	17068	83.13
2	2236	10.89
3	669	3.26
4	244	1.19
5	90	7.73
> 5	224	1.09

These statistics reveal an interesting trend, specifically that most authors have published only one paper, which means that they are unlikely to have a large impact on the co-author network. They also tell us that the 1% of authors have been the most productive.

However, it is worth noting that because of the flaws in the dataset, this is not necessarily reflective of scientific output from an individual. As an example, 27 papers have been attributed just to Nature magazine.

These statistics will help us obtain subnetworks that offer clearer insights into the collaboration as compared to the main network a very large chunk of which is authors that have authored just a single paper.

3 Network Analysis

We obtained four different networks from the data we had. They are listed below:

- The main network with 5163 papers and 20531.
- A network excluding authors who have written only one paper, papers with only 1 author, and authors with degree 1. This has 1623 papers and 3463 authors.
- A network with the largest connected component of the reduced network. This has 1506 authors.
- A network with the top 1% of authors. This has 320 papers and 224 authors.

We start by performing centrality analysis on each of the graphs and obtaining the degree distribution, betweenness, and the relationship between them. The degree can serve to give an idea about the most collaborative researchers and the betweenness can provide us with information about which researchers connect different communities of academics i.e. academics who are collaborating across domains or across universities/teams.

Next, we perform community analysis on the network. We are looking to discover different types of communities on each of the networks. We implement various clustering algorithms to discover communities and use modularity to judge the quality of the communities we have found. Using this, we analyse the capacity of different clustering algorithms to find communities accurately and effectively without encountering the resolution limit.

We use the following clustering techniques in our analysis:

- Multilevel Clustering (Louvain): This method performs an agglomerative clustering on the graph using the modularity as the loss function.
- Leiden Clustering: This method involves iteratively moving nodes to find a partition, and then refining the partitions using a quality function like modularity.
- InfoMap Clustering: This agglomerative method involves iteratively moving each node to a neighbouring node/module and the movement with the largest decrease in map equation is retained.

We have not included all the visualisations for all the graphs owing to space concerns. They are included in the ipynb notebook we have shared.

4 The Networks

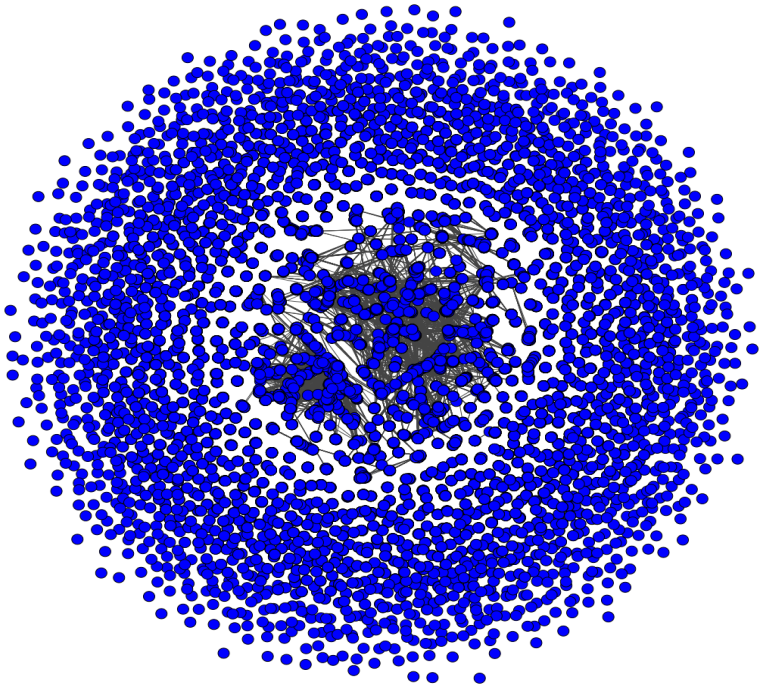
Now, we perform the centrality and community analyses on the four networks we described above and discuss the results:

4.1 The Main Network

We visualise the main network as shown below:

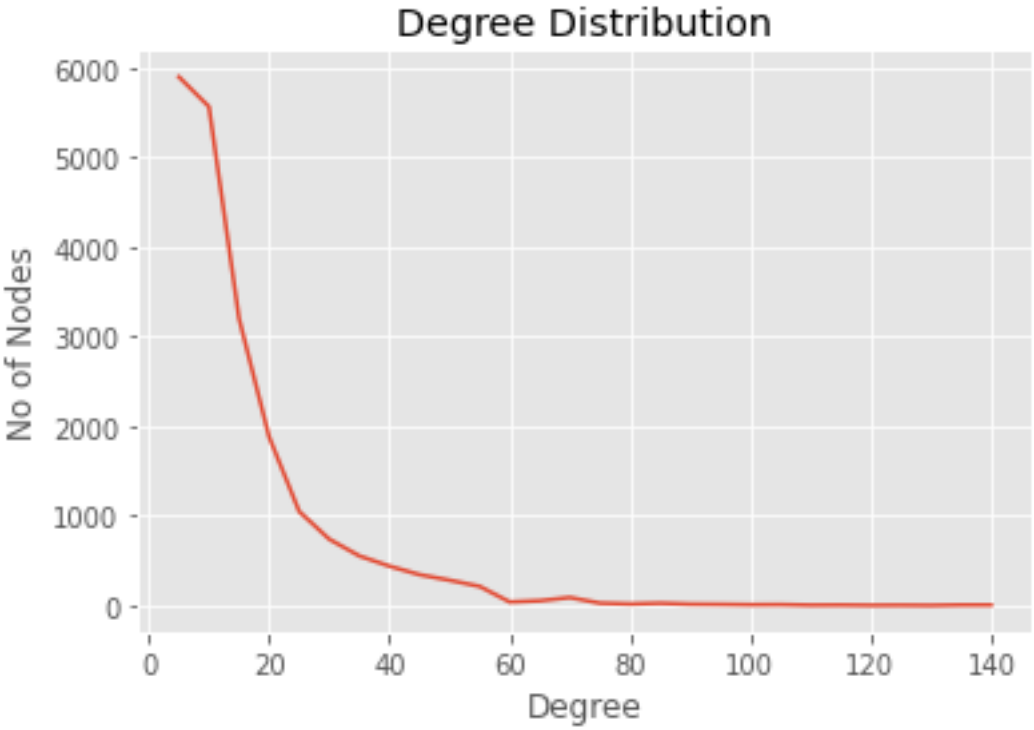
We start by applying the centrality measures to the main network. We get the following information:

We now look at the degree distribution of the network



Degrees		
Minimum	Maximum	Average
0	277	12
Betweenness		
Minimum	Maximum	Average Path Length
0.00	2400155.05	5.16

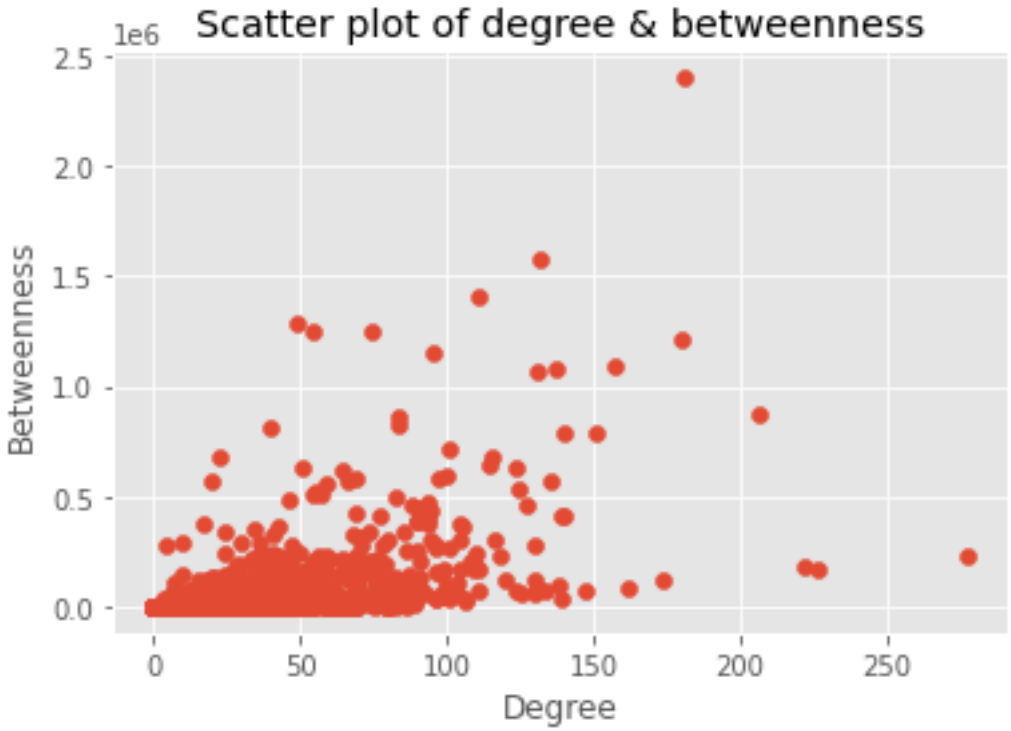
Table 1: The Degree and betweenness for the main graph



We can see that most nodes have a degree of 1, which is not surprising as we know that

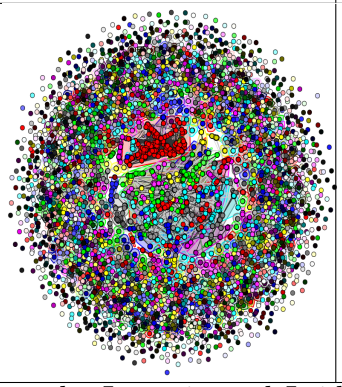
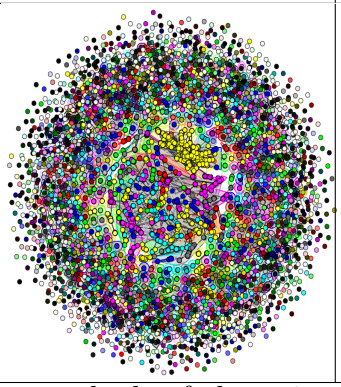
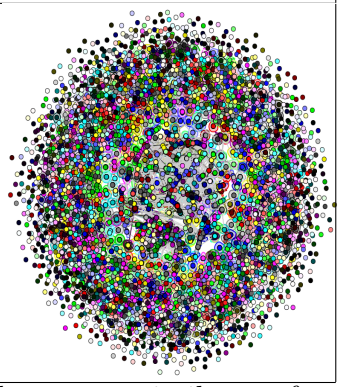
most authors have only published a single paper, and they might well have published it alone.

Next, we look at the relationship between the degree and betweenness and plot the values for each node on the scatter plot below:



Here, we can see that most nodes have a low degree and betweenness while some outliers indicate high collaboration and scientists connecting large communities.

Next we perform community analysis on the network. We use the well-known multilevel, Leiden, and infomap techniques to obtain communities, compare them on the basis of modularity, the quality of the communities as judged by multiple plots, and the number of communities:

	Multilevel	Leiden	Infomap
Communities	2686	2688	3194
Modularity	0.95	0.95	0.89
DrL plot			

We can see that the Louvain and Leiden methods of clustering have near-similar performance. They have almost the same number of communities and they have the same modularity. This is interesting as the Leiden paper [2] claims to achieve better resolution performance than the Louvain algorithm. However, this is also because we are using modularity as the loss

metric in the Leiden algorithm. If we use the MAP metric, as we do in the InfoMap algorithm, we get better resolution but a lower modularity.

4.2 The Reduced Network

We visualise the reduced network as shown in fig 3

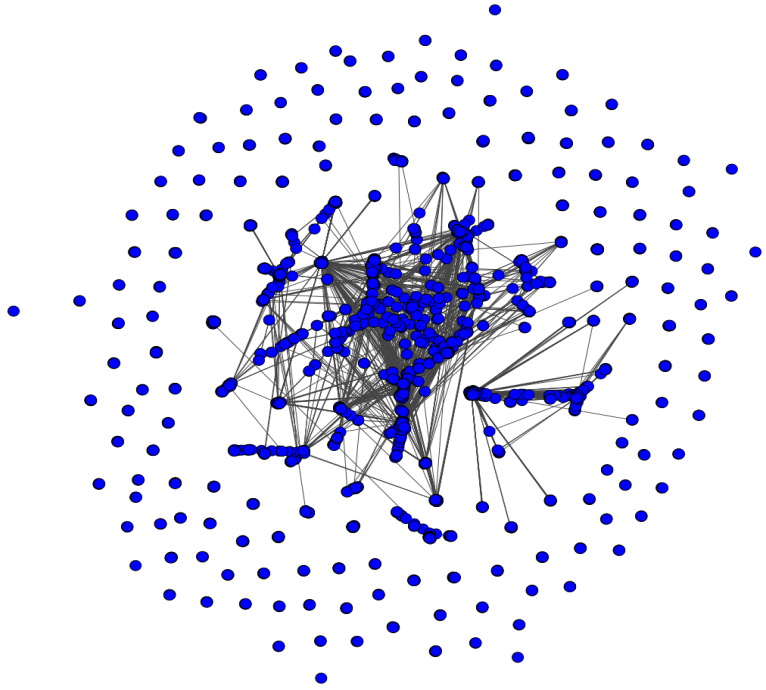


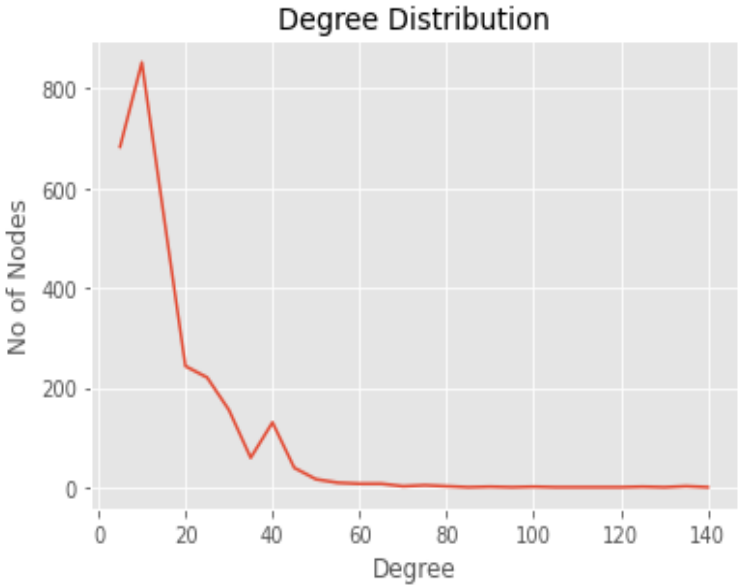
Figure 3: The Reduced network

As before, we apply the centrality measures to the reduced network. We get the information in table 2:

We now look at the degree distribution of the network:

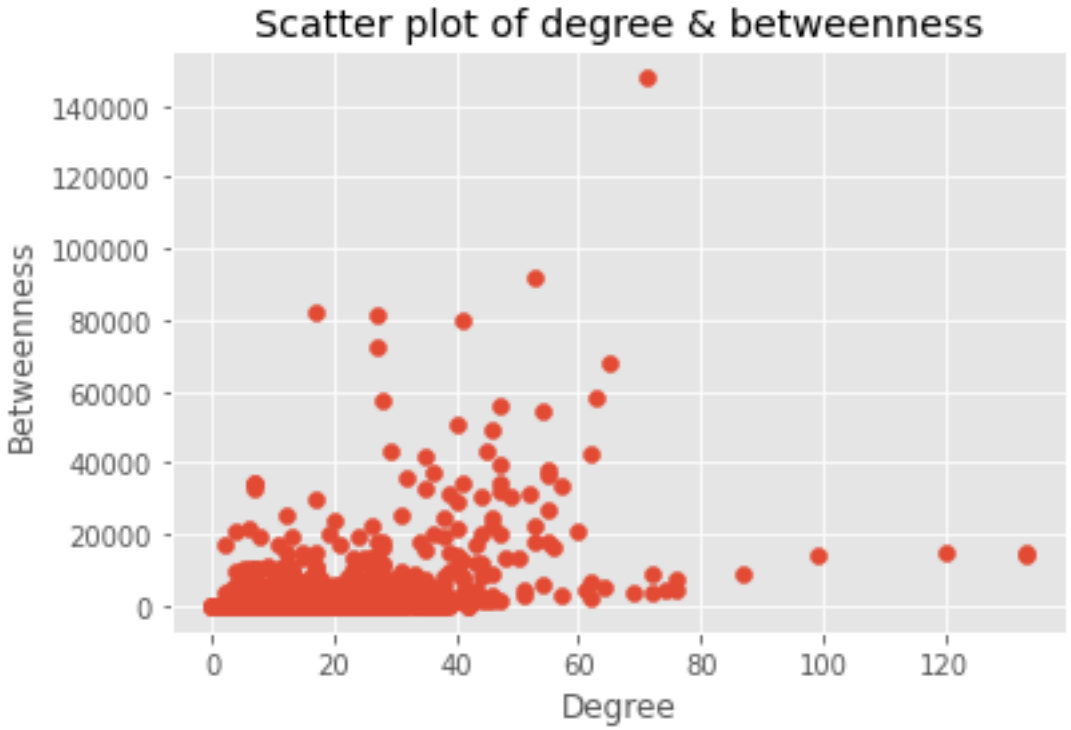
Degrees		
Minimum	Maximum	Average
0	133	12.9
Betweenness		
Minimum	Maximum	Average Path Length
0.00	147848	4.55

Table 2: The Degree and betweenness for the reduced graph



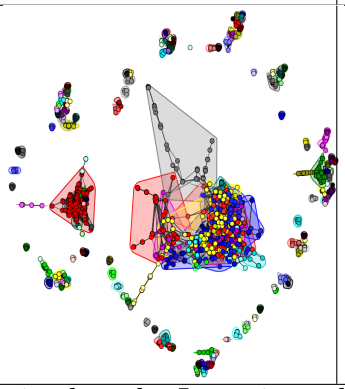
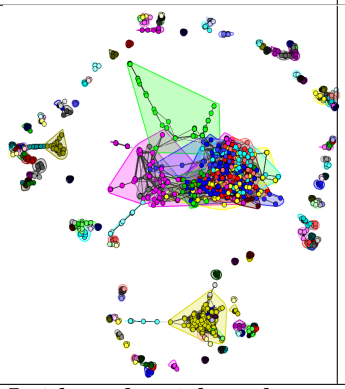
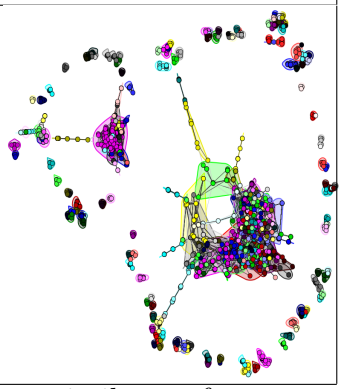
We can see that most nodes have a degree greater than 1, as we have removed most of the authors with a single paper.

Next, we look at the relationship between the degree and betweenness and plot the values for each node on the scatter plot below:



We can see that there are not a lot of visible differences from the main graph, suggesting that removing the authors with one paper hasn't caused a large difference in the relationship between the degree and betweenness.

We analyse the communities for the reduced network now:

	Multilevel	Leiden	Infomap
Communities	178	179	295
Modularity	0.88	0.88	0.84
MDS plot			

We can see again that the Louvain and Leiden algorithms have near-similar performances in terms of number of communities and modularity, and InfoMap performs better in terms of finding communities.

4.3 The Largest Connected Component

We visualise the largest connected component of the reduced network as shown in fig 4

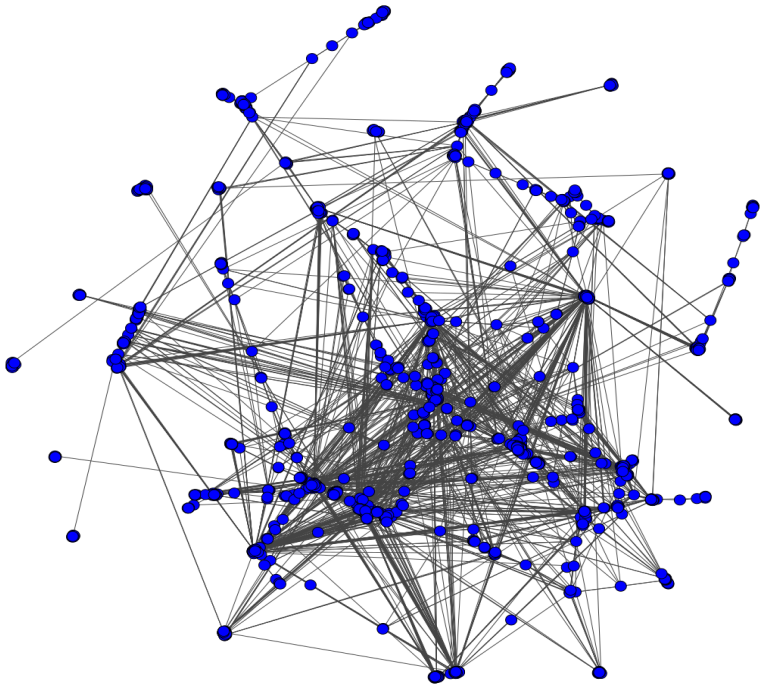


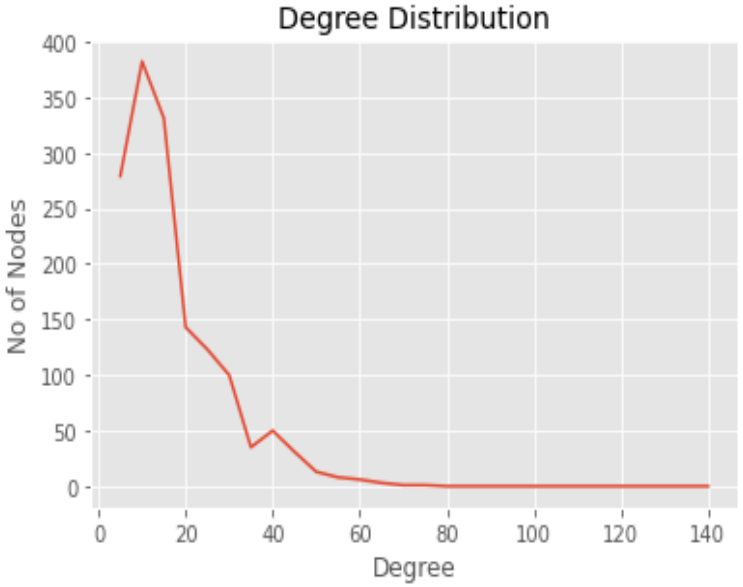
Figure 4: The Largest Connected Component

As before, we apply the centrality measures to the largest connected component. We get the information in table 3:

We look at the degree distribution of the network:

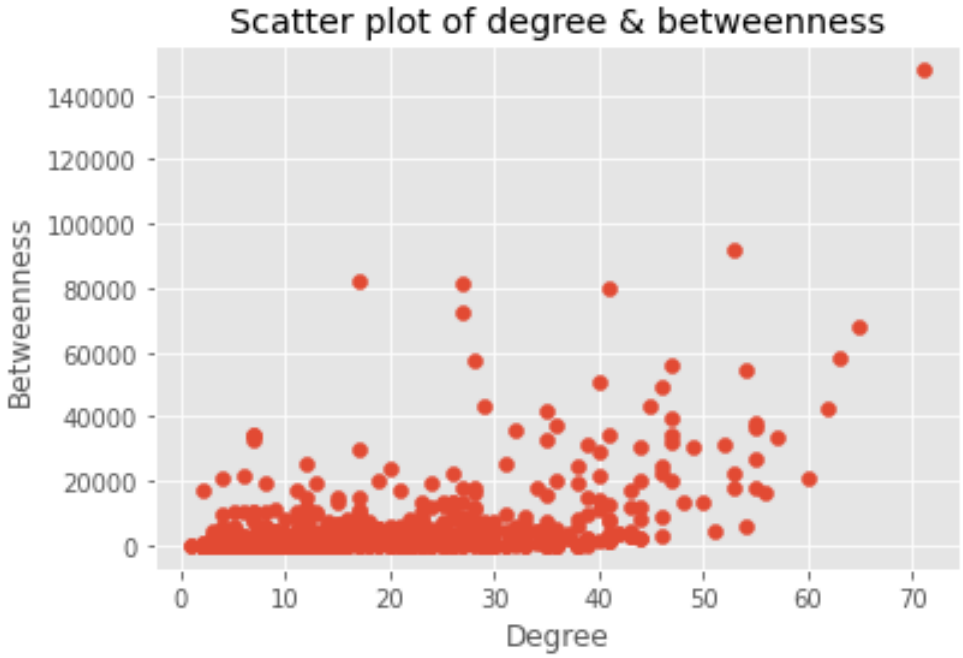
Degrees		
Minimum	Maximum	Average
1	71	14
Betweenness		
Minimum	Maximum	Average Path Length
0.00	147848	4.73

Table 3: The Degree and betweenness for the largest connected component



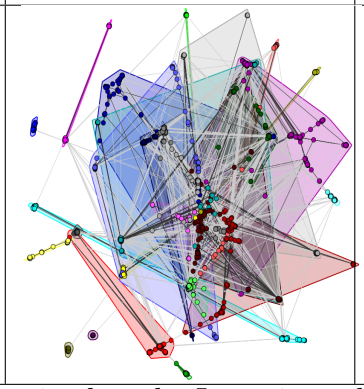
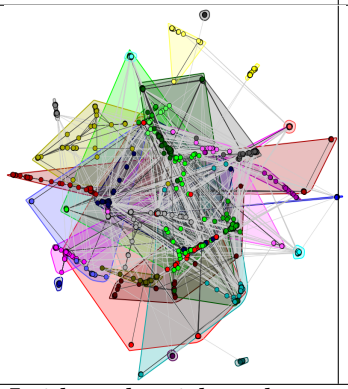
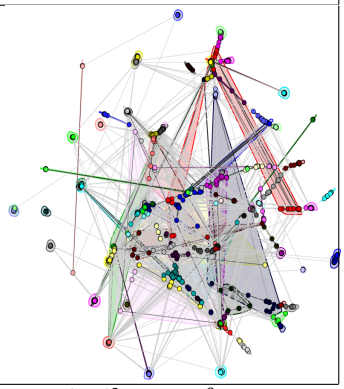
We can see that most nodes have a degree greater than 1, as all of the components are connected.

Next, we look at the relationship between the degree and betweenness and plot the values for each node on the scatter plot below:



We can see that there are a lot of visible differences from the main graph, suggesting that this component does behave differently in terms of degree and betweenness compared to the previous networks. Also, the node with the maximum betweenness is the same as the reduced network, and it is also the same as the node with the highest degree in this graph, meaning that it is a node connecting large communities and is also a node with a large number of connections.

We analyse the communities for this graph now:

	Multilevel	Leiden	Infomap
Communities	26	28	113
Modularity	0.82	0.83	0.78
DrL plot			

We can see again that the Louvain and Leiden algorithms have near-similar performances in terms of number of communities and modularity, and InfoMap performs better in terms of finding communities.

4.4 The Network with the top 1% of authors

We visualise the network with the top 1% of authors of the main network as shown in fig 5. We have visualised this network with a separate technique that visualises the size of a node based on their degree.

As before, we apply the centrality measures to this network. We get the information in Table 4:

We look at the degree distribution of the network:

Degrees		
Minimum	Maximum	Average
0	33	6.7
Betweenness		
Minimum	Maximum	Average Path Length
0.00	2608	3.86

Table 4: The Degree and betweenness for the network with the top 1% of authors

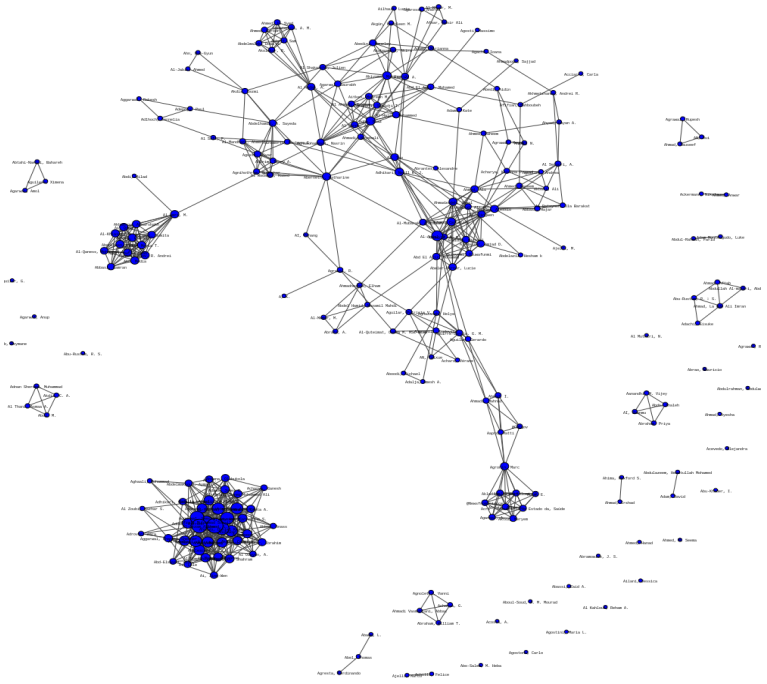
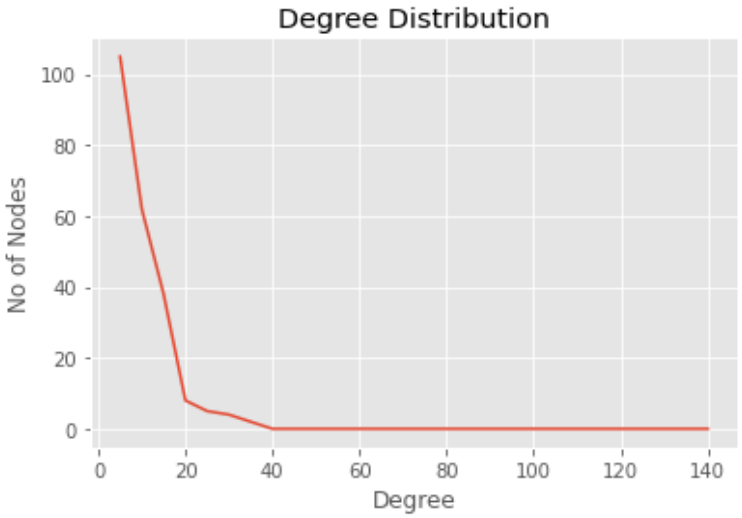
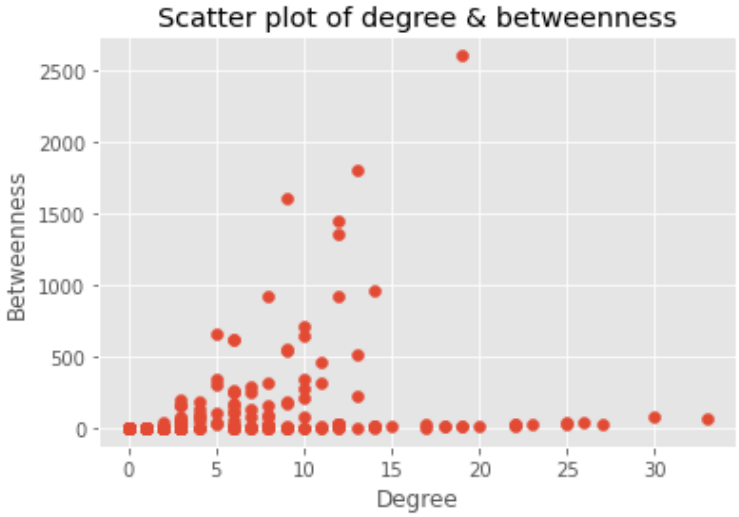


Figure 5: The Network with the top 1% of authors



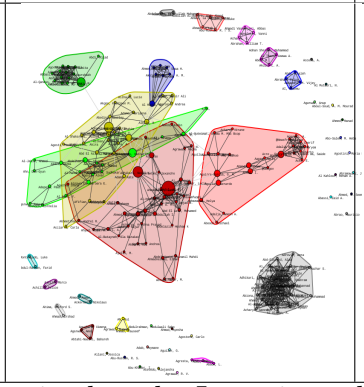
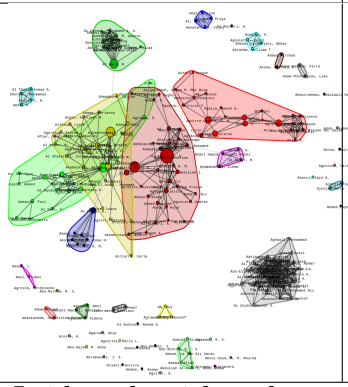
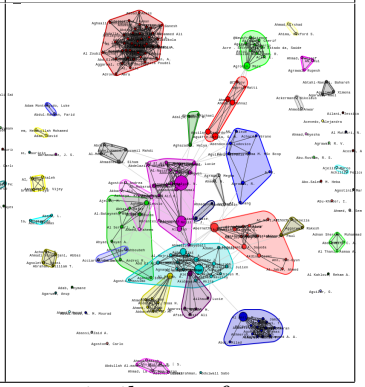
We can see that the degree plot is closer to the original network compared to the reduced networks

Next, we look at the relationship between the degree and betweenness and plot the values for each node on the scatter plot below:



We can see that there are a lot of visible differences from the reduced graph, suggesting that this component does behave differently in terms of degree and betweenness compared to the previous networks.

We analyse the communities for this graph now. We have again used a different visualisation technique here that correlates a node's size to its betweenness.

	Multilevel	Leiden	Infomap
Communities	41	42	52
Modularity	0.73	0.73	0.70
FR Plot			

We can see again that the Louvain and Leiden algorithms have near-similar performances in terms of number of communities and modularity, and InfoMap performs better in terms of finding communities.

5 Qualitative Analysis

We analyse some qualitative aspects of the top 1% author network below. We talk about the authors with the highest betweenness and show what this reflects about them:

To get some insights about the authors, we looked into their profiles.

- Abernethy, Katharine: Senior Research Fellow at University of Stirling
- Adhikari, Neill K. J.: Associate scientist at Sunnybrook Health Sciences Centre
- Aghamohammadi, Nasrin: Associate professor of Environmental and Public Health Leader at University of Malaya

Author	Betweenness	Cluster Author belongs to
Abernethy, Katharine	1357.43	0
Adhikari, Neill K. J.	1809.46	0
Aghamohammadi, Nasrin	1614.61	1
Al Amri, M.	1455.00	2
Al-Ahdal, M. N.	2607.53	1

Table 5: Betweenness qualitative Analysis

- Al Amri, M.: Research Associate at Cardiff University
- Al-Ahdal, M. N.: Adjunct Principal Scientist at King Faisal Specialist Hospital and Research Centre

From this information, we see that there is collaboration between authors from different institutions and countries, for example Universities from UK, Canada, Malaysia and United Arab Emirates, and that these authors are the ones that connect these communities. As an example, Dr. Al-Ahdal connects large researcher communities and thereby gets the highest betweenness score.

Now we look into the authors with highest degrees.

Author	Degree	Cluster Author belongs to
Afzal, Muhammad Sohail	25	0
Aggarwal, Neeraj	26	0
Ahorsu, Daniel Kwasi	27	1
Ailhaud, L.	33	2
Ajlan, Amr M.	25	1
Al-Azzam, Sayer	30	3

Table 6: Degree qualitative Analysis

On implementing the Louvain clustering algorithm on this network, we can identify the below structures:

- Singleton communities of 1 registered author. These are authors with many publications, whose co-authors were removed during pre-processing.
- Small communities of 3-5 authors, disconnected by the rest of the communities. They can be small research teams, that have contributed lots of papers in a small amount of time
- Small communities with 1 author with high betweenness. These are communities with authors connected to each other, and a single author with connections outside of the community.
- Large communities with high degree and small betweenness. These are clusters where there is lots of collaboration between the authors of the same community of Universities or research centers, however not connection with other communities.
- Large connected communities, with many authors and several connections between communities

6 Summary and Conclusion

In summary, we provide a summary of all the analyses for the graphs.

Network	Size		Degree			Betweenness		Avg Path
	Nodes	Links	Min	Max	Avg	Min	Max	
Full Network	20531	132394	0	277	12.9	0.0	2400155	5.16
Reduced Network	2980	19261	0	133	12.9	0.0	147848	4.55
Largest Component	1506	10571	1	71	14.0	0.0	147848	4.73
Top 1% Authors	224	752	0	33	6.7	0.0	2608	3.86

Table 7: Network Centrality Summary

Network	Louvain		Leiden		Infomap	
	Clusters	Modularity	Clusters	Modularity	Clusters	Modularity
Full Network	2686	0.95	2688	0.95	3194	0.89
Reduced Network	178	0.88	179	0.88	295	0.84
Largest Component	26	0.82	28	0.83	113	0.78
Top 1% Authors	41	0.73	42	0.73	52	0.70

Table 8: Clustering Summary

In conclusion, we analysed various centrality measures and clustering techniques on various versions of a Covid-19 Co-author Colloboration graph. We can see that removing authors who hadn't collaborated gives us a richer picture of the communities, and how different clustering techniques provide different clusters. We can see clearly across the four networks that the Louvain and Leiden techniques (while using modularity as the quality metric) have near-similar performances in terms of the number of communities and modularity. The infomap algorithm provides better resolution than the other two algorithms. This can be ascribed to the map quality metric as it overcomes some of the problems of the modularity metric.

7 Contributions

We both did 50% of the work each. A lot of our work was done together, so it is difficult to divide it into easily separable chunks. The ideation was done together, so were most of the tasks. The approximate separation of work has been listed in table 9:

Elina	Siddharth
EDA	Centrality Analysis
Relevant Networks Definition	Clustering Analysis
Centrality Analysis	Refactoring
Clustering Analysis	Small parts of the EDA
Qualitative Analysis	Report

Table 9: Contributions

References

[1] Miloš Savić, Mirjana Ivanović, and Lakhmi C. Jain. *Co-authorship Networks: An Introduction*, pages 179–192. Springer International Publishing, Cham, 2019.

[2] Vincent Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *arXiv preprint arXiv:1810.08473*, 2018.

[3] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset. *ArXiv*, abs/2004.10706, 2020.