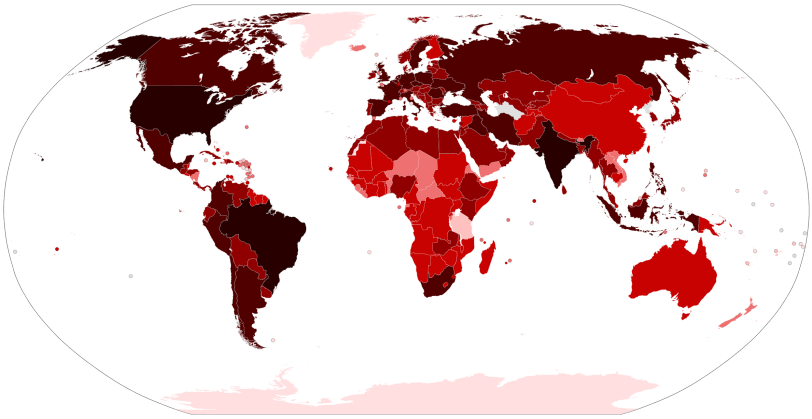
Complex networks - project

International flights during covid 19



Submitted by:

Olga Soldatenko 342480308

Yosef Batash 061106605

Introduction

It is impossible to imagine the modern world without international flights. Not only individuals depend on them, but also big and small businesses, post services, etc. Due to Covid19 pandemic the big part of air travel was restricted which has crippled the supply chains, led to huge money losses and entrapped people outside of their home countries.
Such a complex and difficult situation is important for understanding in detail, which makes the dynamic of the flight network during the time of crisis an interesting subject for research.
The purpose of our work was to explore the international flights' network during the covid19 in order to try and analyze the impact of the pandemic on international transportation. Analysis of the network could help to understand the changes in traffic and behaviours of certain 'players' (countries).
In order to enrich our outlook we have cross referenced the flights data with the covid19 database in search of insights.
All of the code files and gephi graphs are available at the git repository.

Data

Background

Our biggest data set, worldwide flight data between the years 2019 - 2021 (approximately ~ 4.2GB), retrieved from the open sky network.
The OpenSky Network is a non-profit association based in Switzerland. It was set up as a research project by several universities and government entities with the goal to improve the security, reliability and efficiency of the airspace. Its main function is to collect, process and store air traffic control data and provide open access to this data to the public.

Data sources and ETL

In our data exploration phase we aggregated the flight data set in multiple different windows of time (months, weeks) and chosen level of detail (in resolution of airport or country).
Here is a short description of our ETL/data pipelines process:

- Flight data was downloaded with an automated bash script (00_download_flight_data.sh) from this site (schema description also available there).
- After we did exploration on the data we decided that we need to reduce our data set slice dice and aggregate it so we can work with it (gephi and compute power limitations...), from the raw data we created a few different aggregations:
    - Flights_each_day_week_14_data - for each year we aggregate by date, airplane_code, origin, destination and count all flights with origin and destination not null for week number 14.
    - Source_target_count_week_14_per_year - from the aggregate data set we did another aggregation by year, airport source and destination and sum flights count.

- 03_create_dim_countries- to enrich our nodes data we created a country dimension - with attributes like - iso2 code,iso3 code, country, continent, latitude, longitude.

- 04_countries_and_covid_19_week_14_data - for each country we added Covid19 metrics such as - country population, total cases, and cases per population ratio.
- 05_create_airports_nodes_per_year - for each of our network (flights for week 14 in year 2019 | 2020 |2021) we created the proper nodes data set.
- 06_international_flights_each_day_week_14_data - In this data flow we filtered out all non international flights for the daily data set.
- 07_international_source_target_count_week_14_per_year - In this data flow we filtered out all non international flights for the daily data set(aggregated by year).
- 08_international_country_source_target_count_week_14_per_year - for better visualization we created this data set that aggregates all airports to the airport country (aggregated by year).
- 09_airports_networkx_2019 - in this notebook we use Networkx to investigate our different graphs.
- 10_create_countries_graphs_for_gephi - this notebook used us to investigate the graphs at a high level (by country).
- 11_create_graphs_for_gephi - With this notebook we created graphs for our more grained data set but due to the size we weren't able to visualize it with gephi.

Figure 1 presents the whole pipeline on a high level.

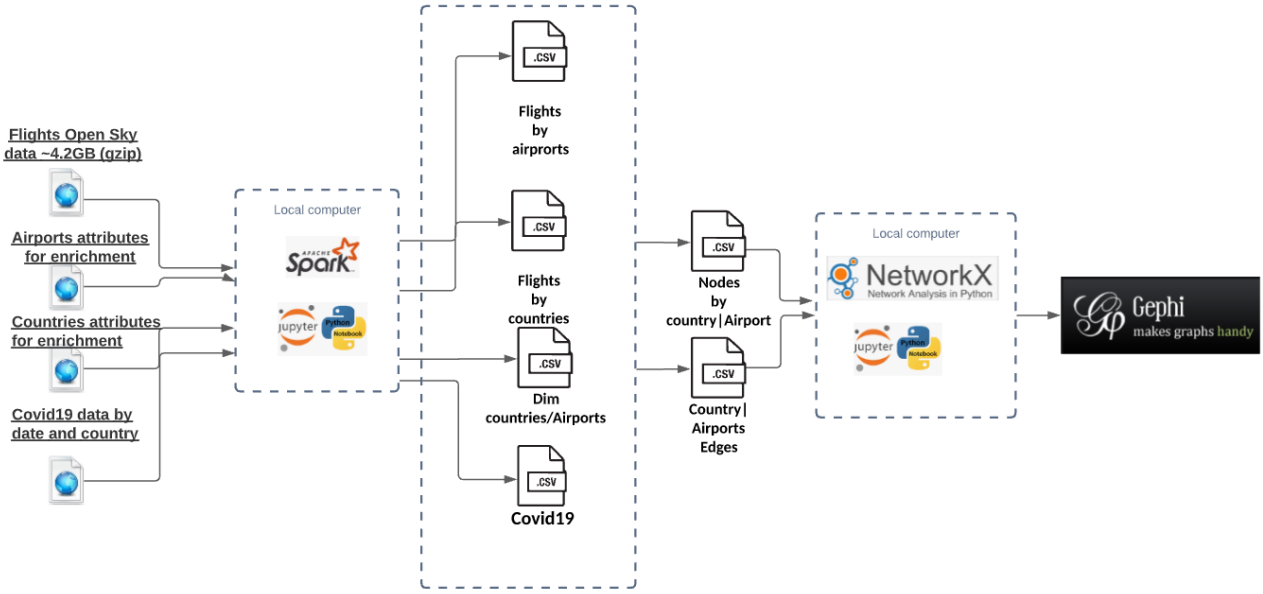


Figure 1: Data pipeline outlook

Network

We have chosen to build a directed weighted graph, where nodes are airports and the edges are flights between them, what is more, the weights correspond to aggregated number of flights between the two nodes. Different windows of aggregation were chosen and examined.
Each node's attributes contained information about the airport's country, type, it's geolocation and percentage of covid cases in the country (for further stages in research).

Firstly, the aggregation by month was performed. It is a very high scale glance on tha data - not very specific and is able to catch only the general trends. We have traced the number of nodes and edges through the whole horison of tracking. The results are presented in **figure 2** and **figure 3**.

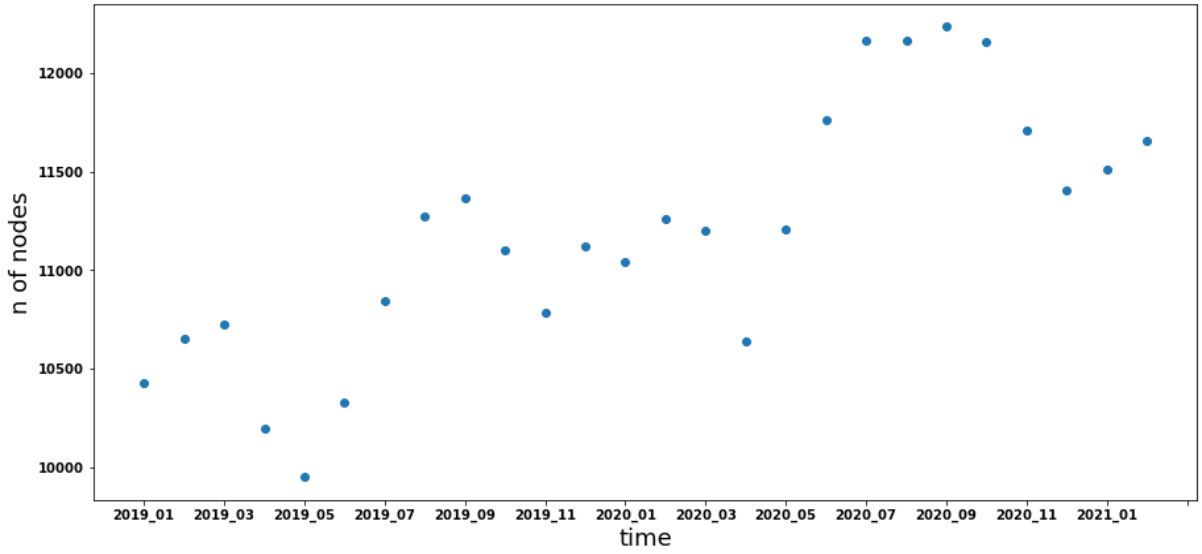


Figure 2: Number of nodes vs the month

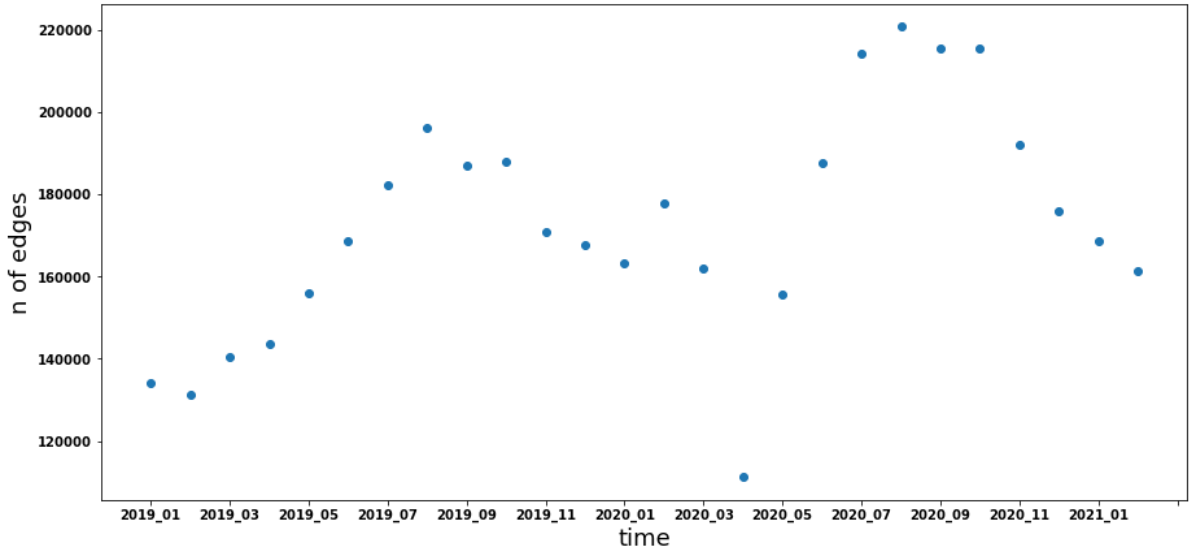


Figure 3: Number of edges vs the month

From the graphs we see the absolute minimum of functioning airports and number of flights in April-May od 2019 with steady growth followed by yet another drop in April of 2020. This behaviour seemed interesting so we have decided to take a closer look at these specific months.
In order to focus our analysis on the details, we have narrowed the aggregation window to a week and took the week 14, that falls on the beginning of April. We have built three networks corresponding to each year in our dataset (2019-2021). **Figure 4** shows node and edge count.

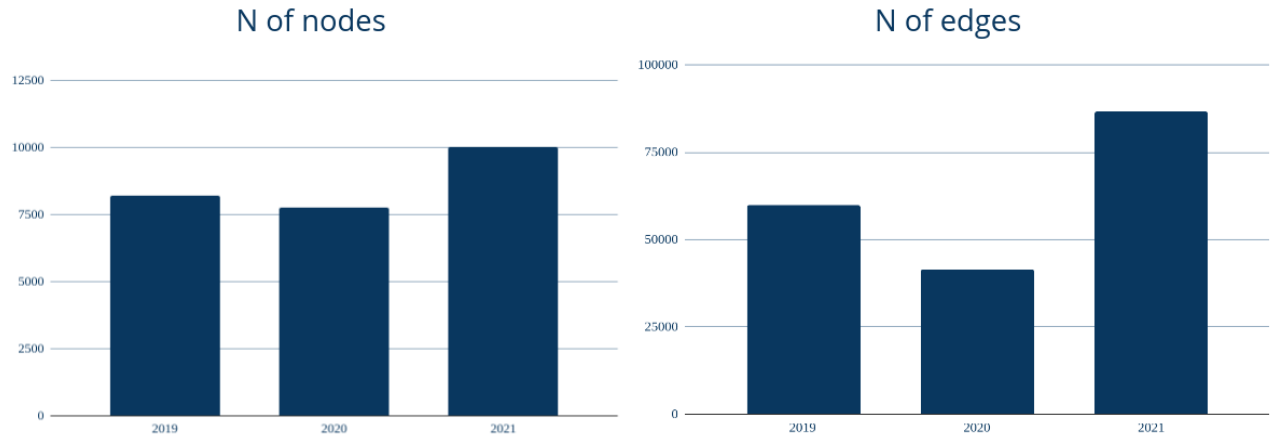


Figure 4: Number of edges vs the year (aggregation by week)

We can see the dramatic difference in the number of flights between the years 2020 and 2021 - 4187 in 2020 against 10264 in 2021.
Due to the existence of airports that do not have incoming flights, the network is not connected - average path length cannot be calculated. This is only logical, as many countries announced a lockdown and only performed rare incoming ‘rescue’ flights in order to bring back their citizens ‘stuck’ abroad.

Degree distribution

Figure 5 shows graphs of in and out degree distributions for all three networks. The networks are clearly scale-free, and there exists a limited number of hubs - highly connected airports. It is the usual state of the airport network. We can see that covid situation has not changed that fact.

Centrality

For the networks centrality measures were calculated - degree and eigenvector centrality.In **figure 6** we see the countries with the most central airports. We see that in 2019 central airports are spread over big economically successful countries such as the United States, Great Britain, France, Germany and the Netherlands. Although we can notice the prevalence of US airports (5-9 place for in degree, 9-12 for out degree, 1-9 for eigenvector). In 2020 and 2021 this dominance is total. US airports catch the top 30 most central places by all methods of calculation. This points to Europe being more cautious and cutting down the traffic as opposed to the US with a more ‘risky’ attitude.

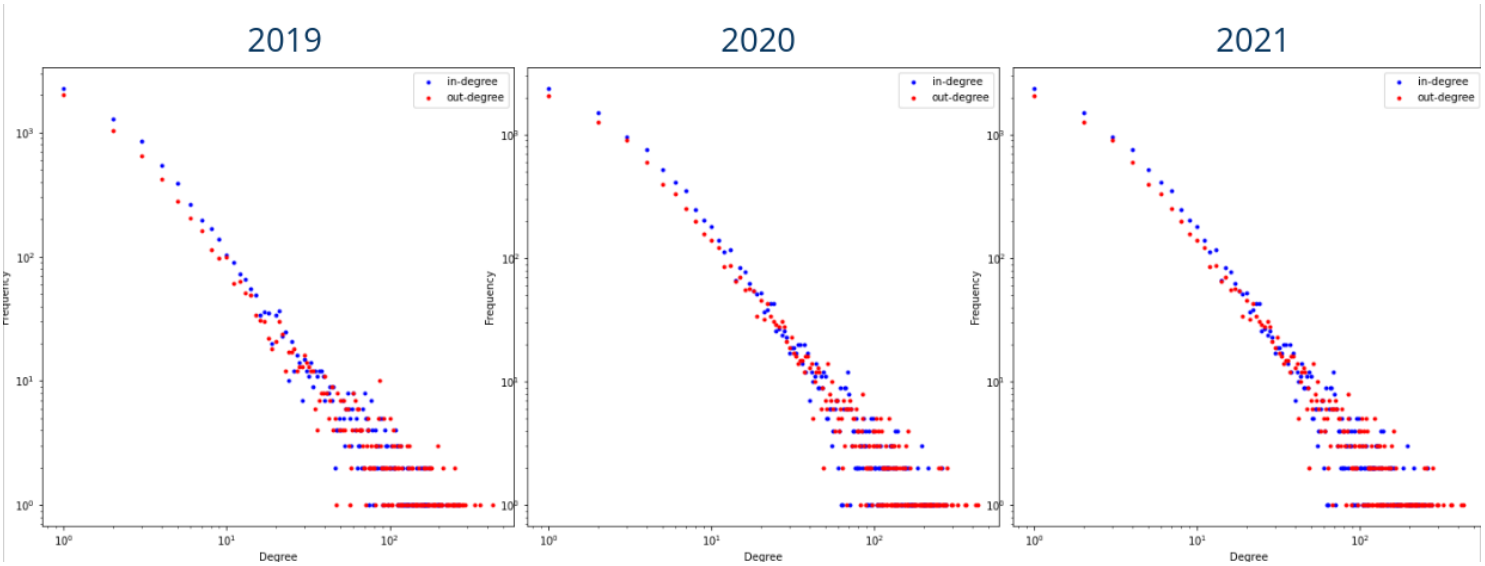


Figure 5: Degree distribution 2019-201 - three scale-free networks

	In degree	Out degree	Eigenvector
2019	**1-3** US	**1-7** US	**1-9** US
	4 NL	**8** NL	**10** GB
	5-9 US	**9-12** US	**11-14** US
	10 DE	**13** DE	**15** DE
	11 FR	**14-15** US	
	12 DE		
	13-15 US		
2020	**Top 30** US	**Top 30** US	**Top 30** US
2021	**Top 30** US	**Top 30** US	**Top 30** US

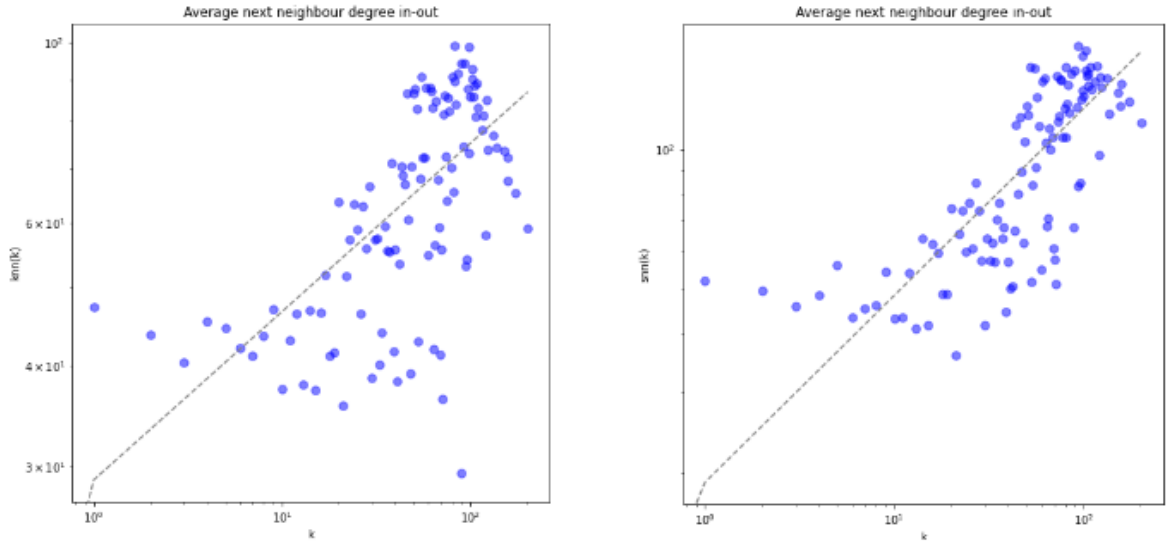
Figure 6: Centrality measures 2019-2021

Clustering coefficient

In order to show the non random character of our network we have compared it to a random one. For this purpose a directed network preserving the number of nodes and edges was built. It's average clustering coefficient is 0.01 against 0.205 in our network.

Degree distribution

Average next in-neighbour out-degree vs out-degree. This graph shows that nodes with higher out-degree have more in-neighbours with high out-degree. The network is assortative. This is also true when weights are taken in consideration (**figure 7b**)- this is even more important as weights represent how busy the destination is.
The assortative nature of the network is present in all other average degree distributions - average out-neighbour in degree, average in-neighbour in-degree and average out-neighbour out-degree (**figures 8-10** in appendix). To summarise - big busy airports are connected to big busy airports with lots of incoming and outcoming flights.
The assortativity coefficient μ for knn(k) is about 0.2-0.17 and for snn(k) is about 0.3-0.4



Figures 7a and 7b: knn(k) and snn(k) in-out degree

Hierarchy

By plotting the average next neighbour clustering coefficient we determine that the network is hierarchical - the neighbours of the hubs are not connected to each other. This makes sense as the airport network is built that way so that bigger airports connect between smaller ones. As we can see in **figure 11** covid19 hasn't changed that.

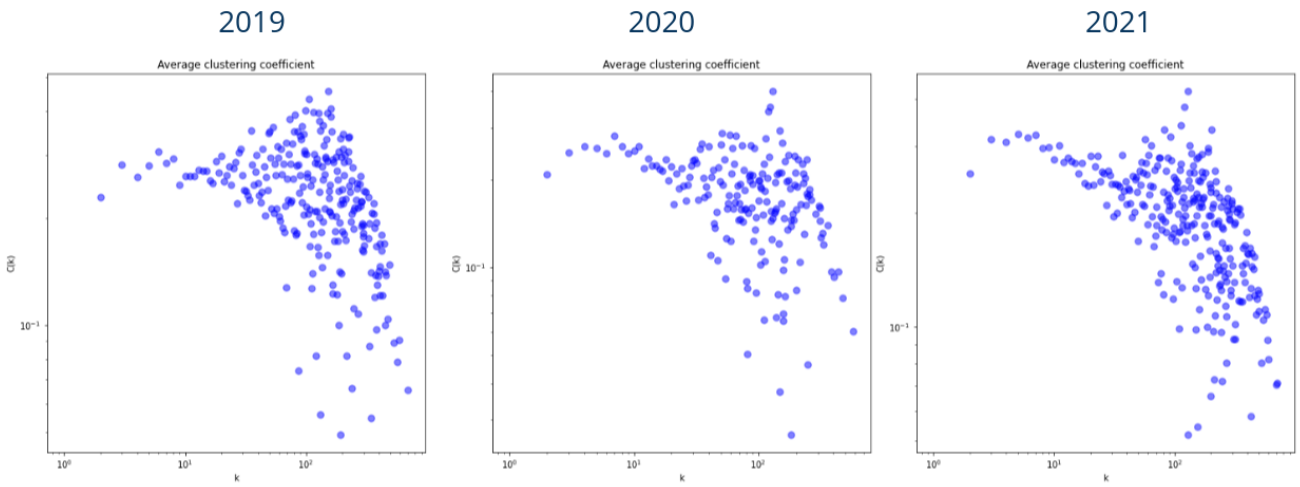


Figure 11: Hierarchy

Aggregation by country

In further work we wanted to focus on country dynamics so we aggregated the data by countries. The graphs built with this aggregation lose their scale-free characteristic (see **figure 12** and **figure 19 a,b** in appendix).

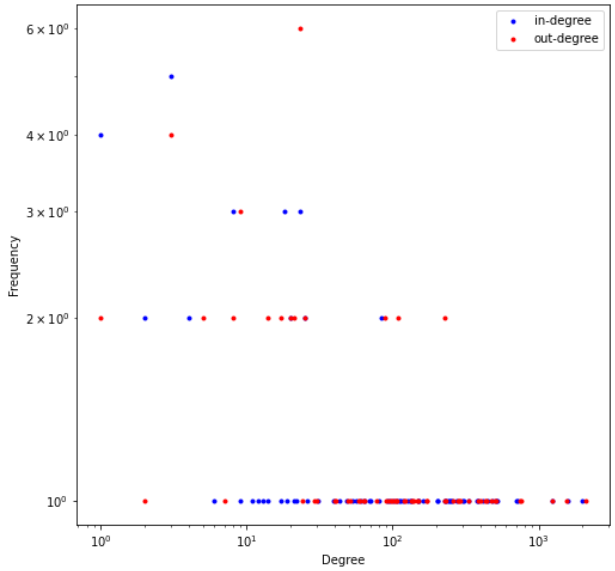


Figure 12: Degree distribution - aggregation by country 2020

Community detection

We have performed community detection by Louvain algorithm and compared it to the partition according to the continents. In figure 13a the graph is partitioned by continents, the node size is according to its degree. In figure 13b the same graph is colloured according to Louvain communities.
We can see that communities more or less correspond to continents, only big and economically active European countries such as UK, Germany, Spain and France are put with the United States, Canada and South America to form kind of an integral continent. Our guess is that it happens due to intensive exchange between the countries as these countries possess the most busy airports that join together the continents.
The similar plots for years 2020 and 2021 are presented in the appendix **figures 20-21ab**. We have noticed that some of the European countries that were clustered with the Americas in 2019 in 2020 'jump' to Europe and then in 2021 go back to Americas. It seems that cross atlantic transportation got weaker.
Another difference in the graph of 2021 is that the fourth community of Africa emerges.

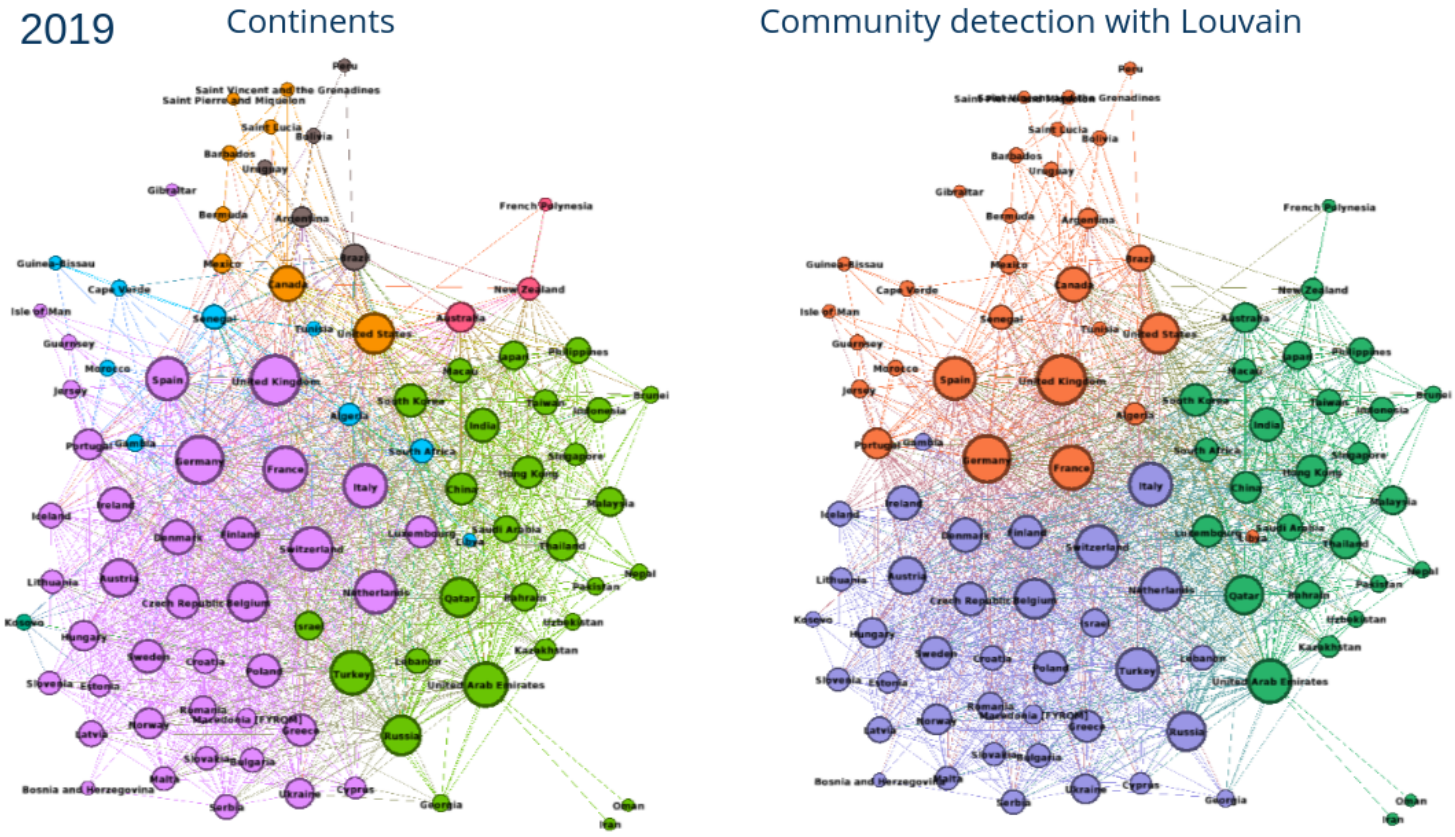


Figure 13 a and b:Continents partition and Community detection in 2019

Visualisation in Gephi

First of all, we have plotted the countries in order to get the overall impression of the network. The visualisations are presented in **figures 13-15**. The graphs are built for week 14 of each year.
The nodes and edges are coloured according to the continents, the edge width corresponds to its weight e.g. the number of flights and the node size represents node's degree.
From the first glance we can notice a dramatic decrease in the number of flights: from 2328 in 2019 to 1433 in 2020 - almost half the volume. In 2021 the situation stabilizes, though not entirely, we have to take in consideration that the network is built for week 14 (April) and the covid situation goes and improves with time. The same observation stands for average weighted degree and is even more prominent: in 2020 on average every country gets less than 0.16 flights than the year before.

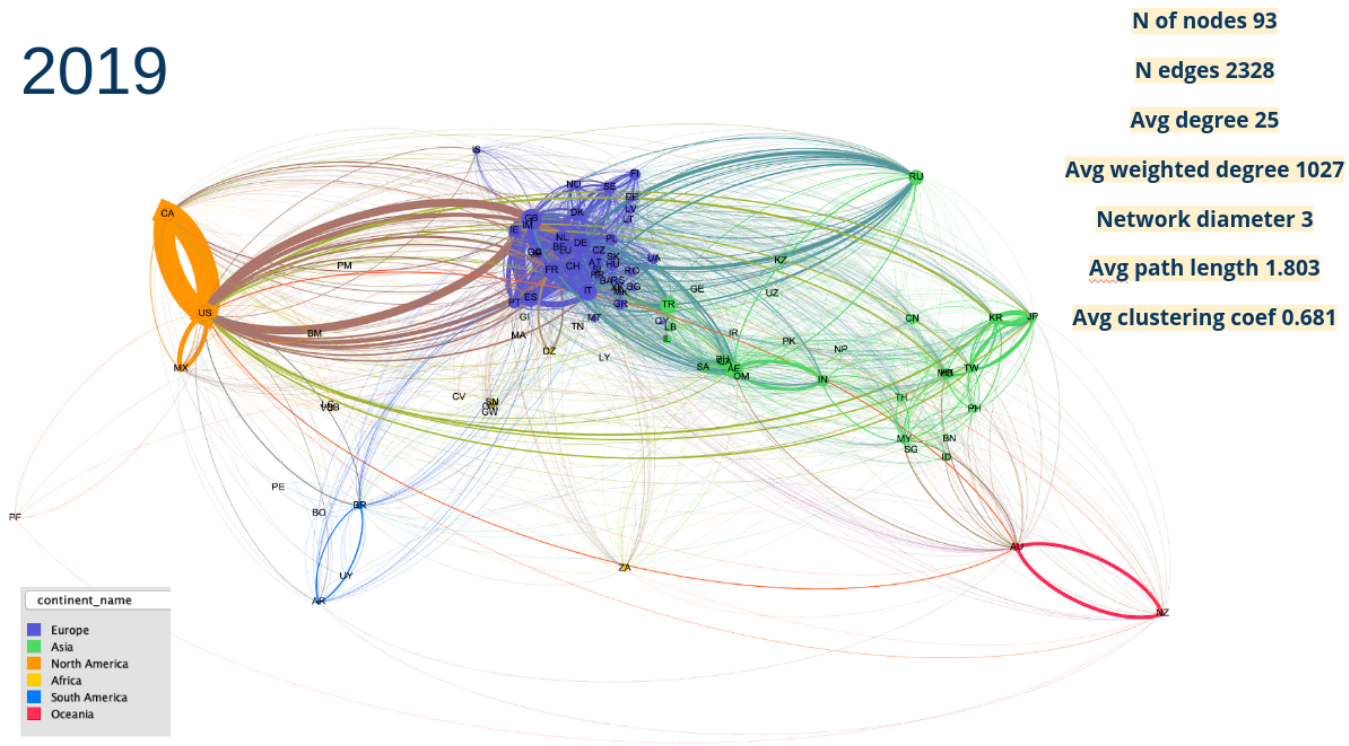


Figure 14: World view - 2019

With the decreasing number of flights network diameter gets bigger in 2020 and 2021 - due to airports and even the whole countries closing up.

2020

N of nodes 88

N edges 1433

Avg degree 16

Avg weighted degree 173.7

Network diameter 4

Avg path length 1.97

Avg clustering coef 0.626

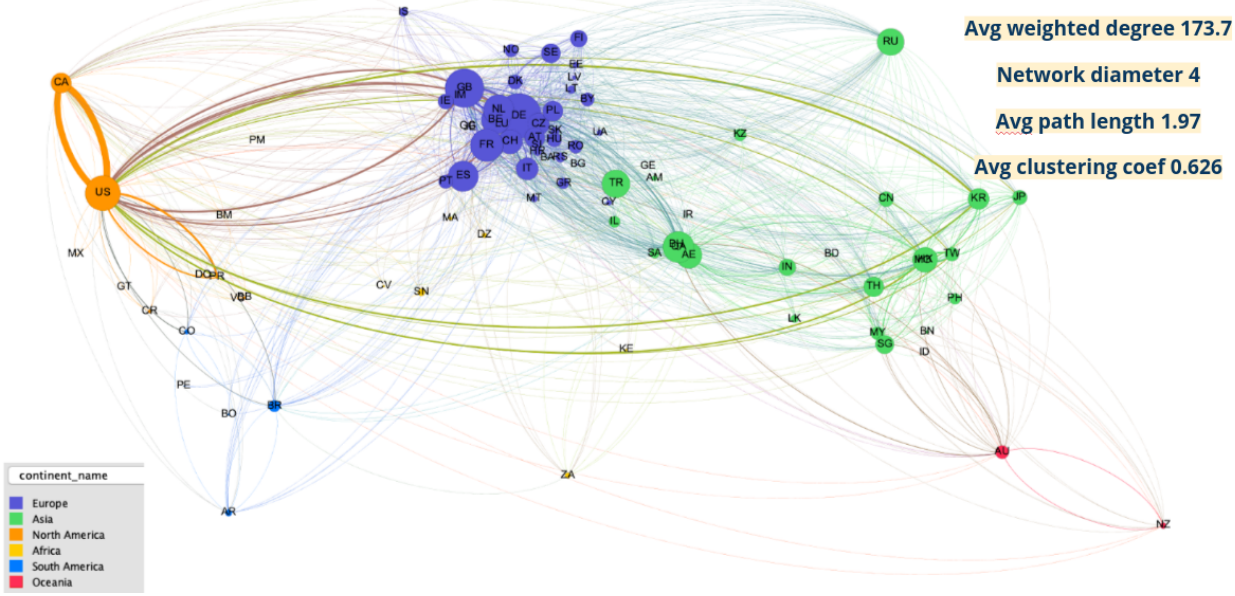


Figure 15: World view - 2020

2021

N of nodes 103

N edges 2202

Avg degree 21.38

Avg weighted degree 411

Network diameter 4

Avg path length 1.963

Avg clustering coef 0.658

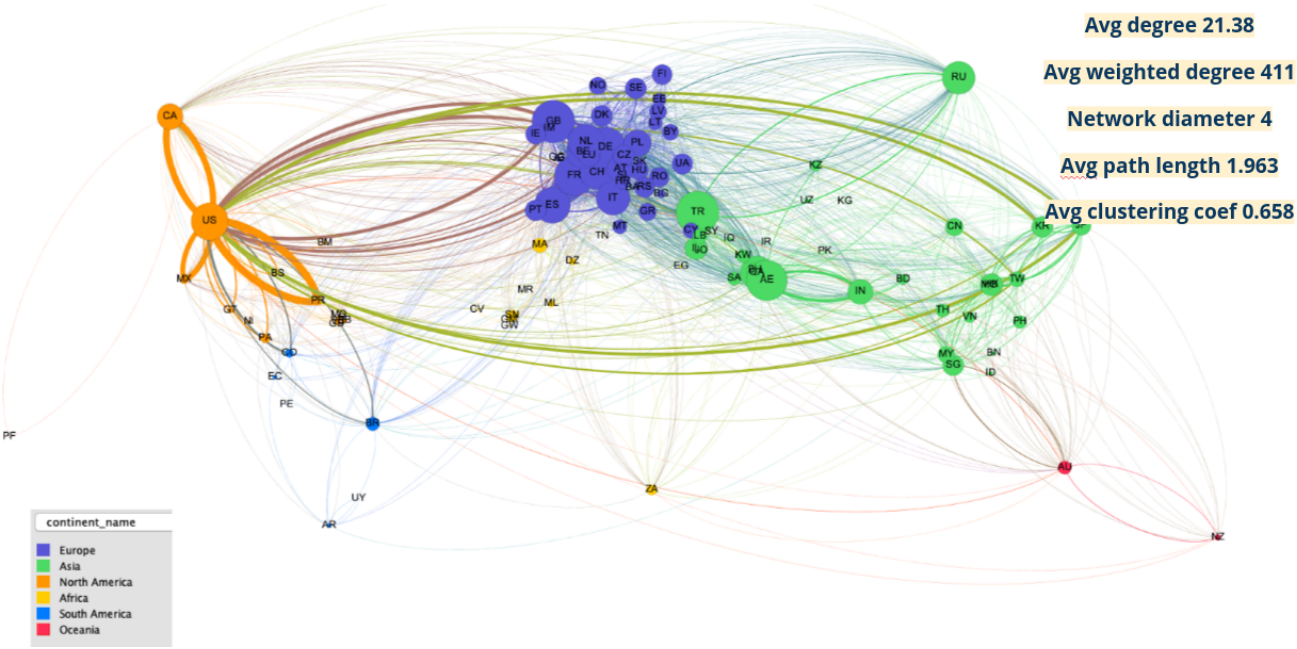


Figure 16: World view - 2021

In **figures 14 and 15** we take a closer look on Europe in terms of covid19 cases rate. The size of the node represents the degree and it's colour - the normalized case rate, the edge colour is in accordance with the number of flights.
We can see an interesting phenomena: in 2020 the nodes are relatively 'light' (not many cases) but the flights merely exist, while in 2020 the epidemiologic situation is much more severe and yet there are relatively many flights. For example, the Czech Republic with 14% covid rate in 2021 and still there are incoming and outcoming flights.
We see that the world adapts, learns how to function in these difficult times, develops new protocols for covid detection and isolation.

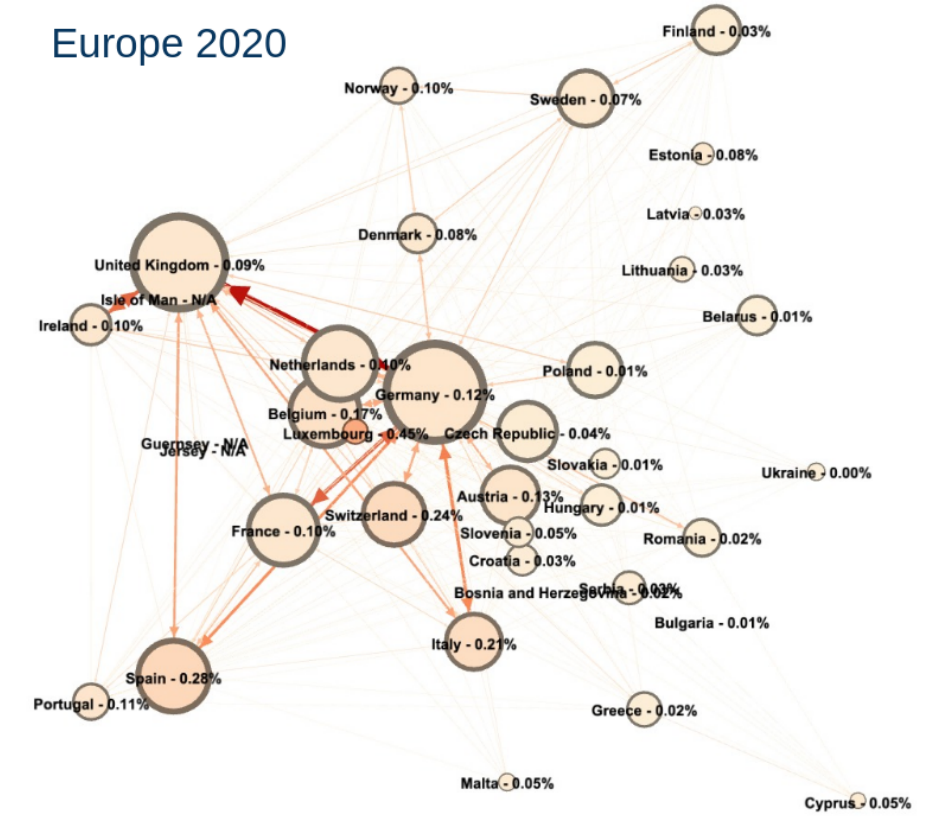


Figure 17: Europe 2020

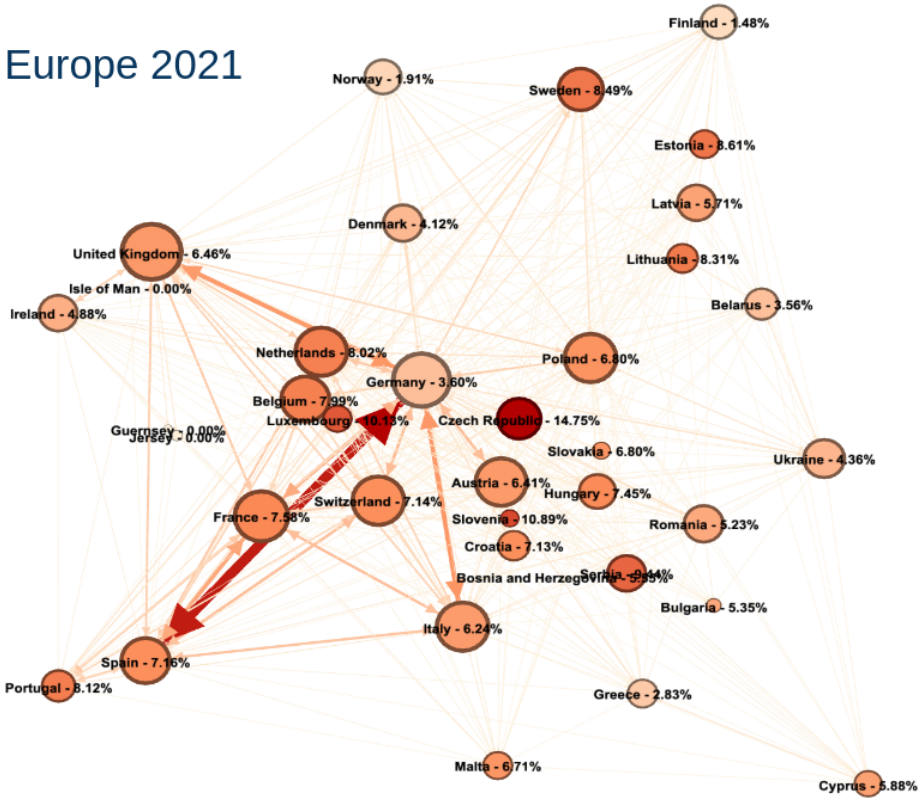


Figure 18: Europe 2021

Summary

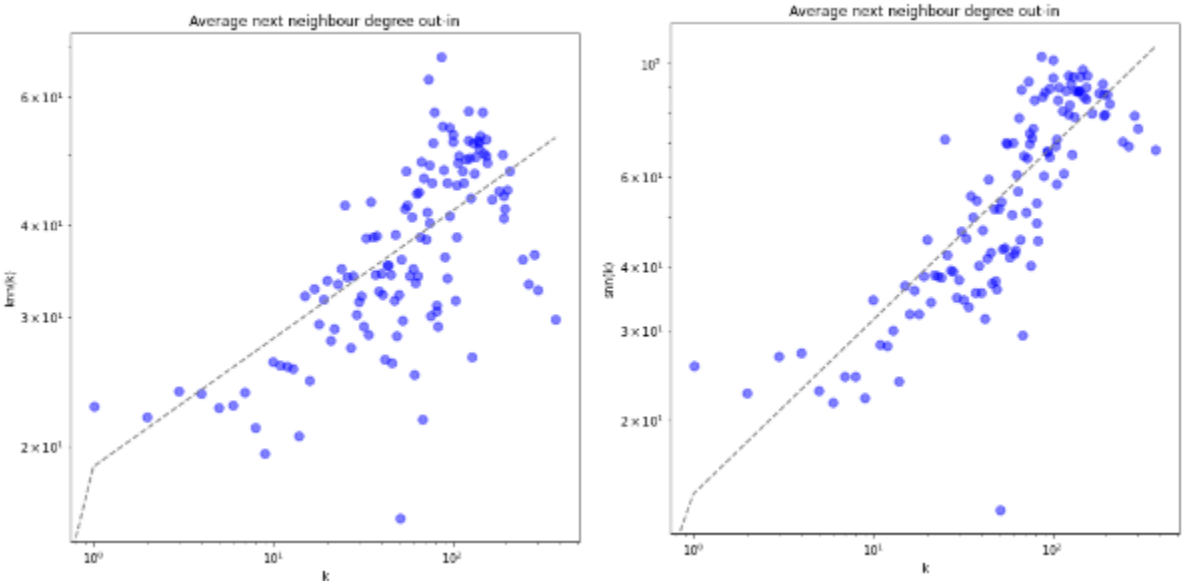
Both airport level and country level networks were examined. We have tracked the international flights network through the last three years, seen in numbers the dramatic fall in its volume, identified the change in central airports - USA becoming the country with the most central airports. We have noticed the change in the covid-related behaviour: from total lockdown to regular regulated flights.

The work doesn't address the actual number of people and goods being transported, this could be done by using the information about specific aircraft types, defining the type of airplane (cargo, passengers, military) and approximating its capacity. This enrichment would open the door to more interesting findings such as change in flight type distribution: maybe there was an increase in cargo flights. Unfortunately, there is no ready open source database of this kind and creating it by ourselves would demand data mining which is out of the scope of this project.

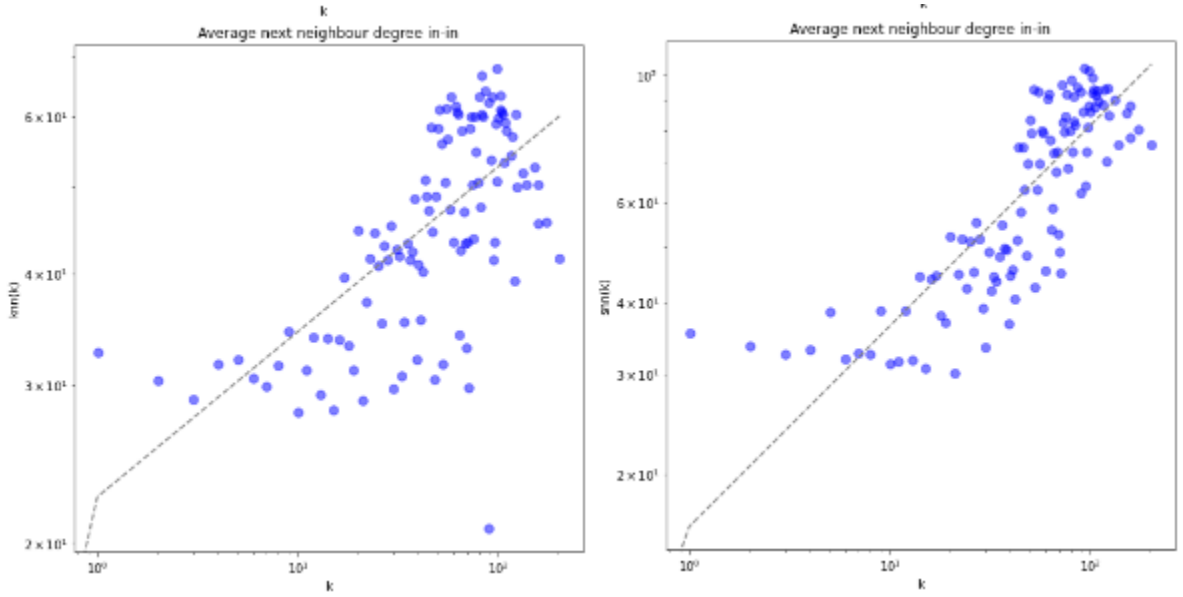
We have attempted to predict the covid case rate in some country two weeks from some time point, based on the network properties at that time point. For each node (country) wa have extracted the parameters of the network such as degree of the node, average degree, average path length, centrality, covid rate of the node at the time point, into a feature vector and made an effort using the regression to predict the covid rate in two weeks. Unfortunately, the results were disappointing.

Maybe by adding more features such as weighted neighbours' covid rate or some other parameters and using other machine learning methods could lead to some success in the prediction.

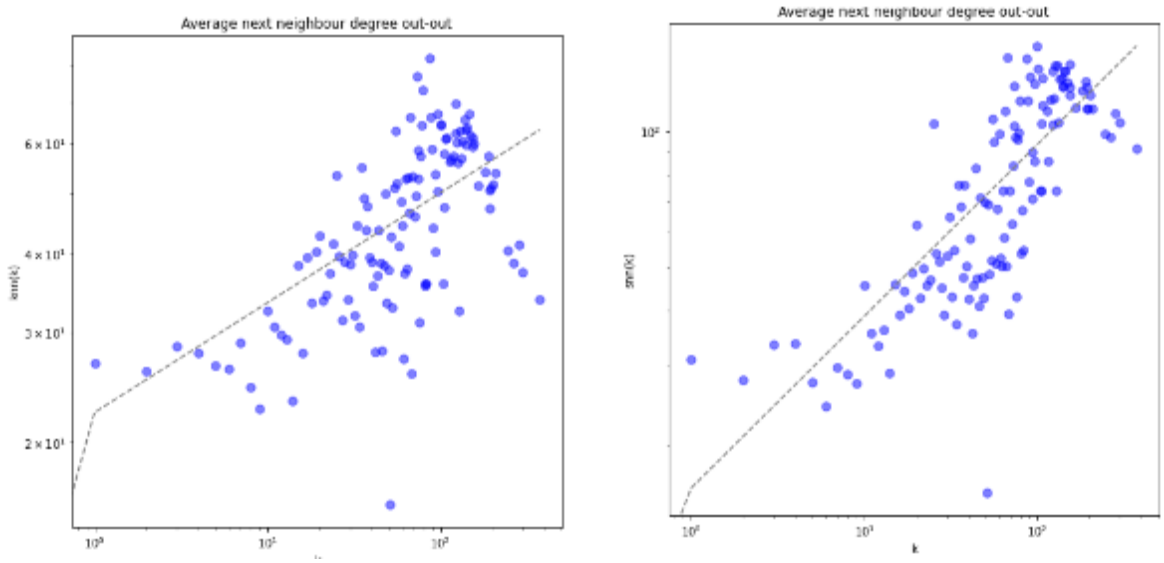
Appendix



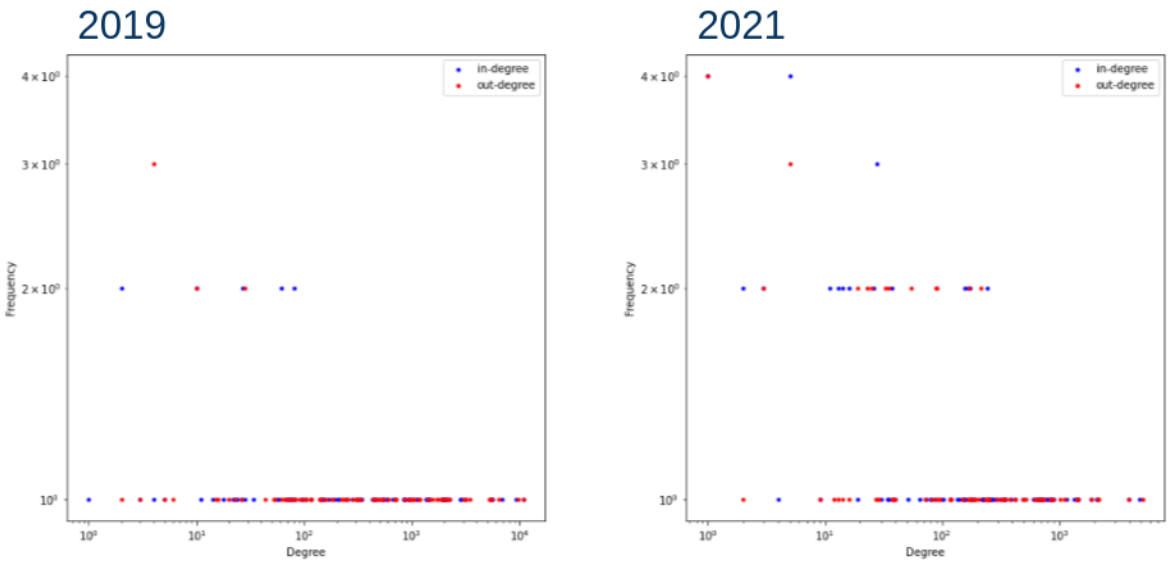
Figures 8a and 8b: knn(k) and snn(k) in-out degree



Figures 9a and 9b: knn(k) and snn(k) in-out degree



Figures 10a and 10b: knn(k) and snn(k) in-out degree



Figures 19a and 19b: Degree distribution for years 2019 and 2021 aggregation by country

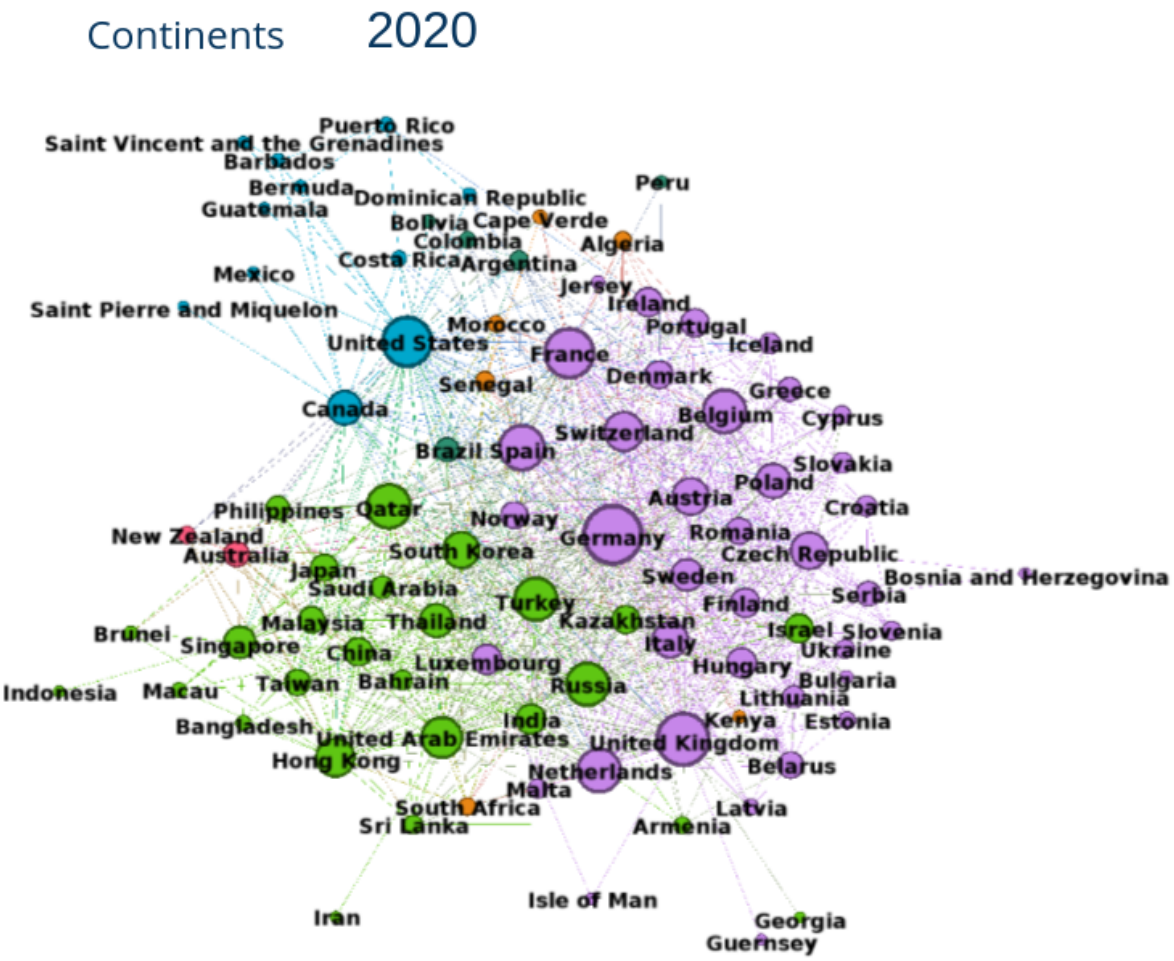


Figure 20a: Continents partition 2020

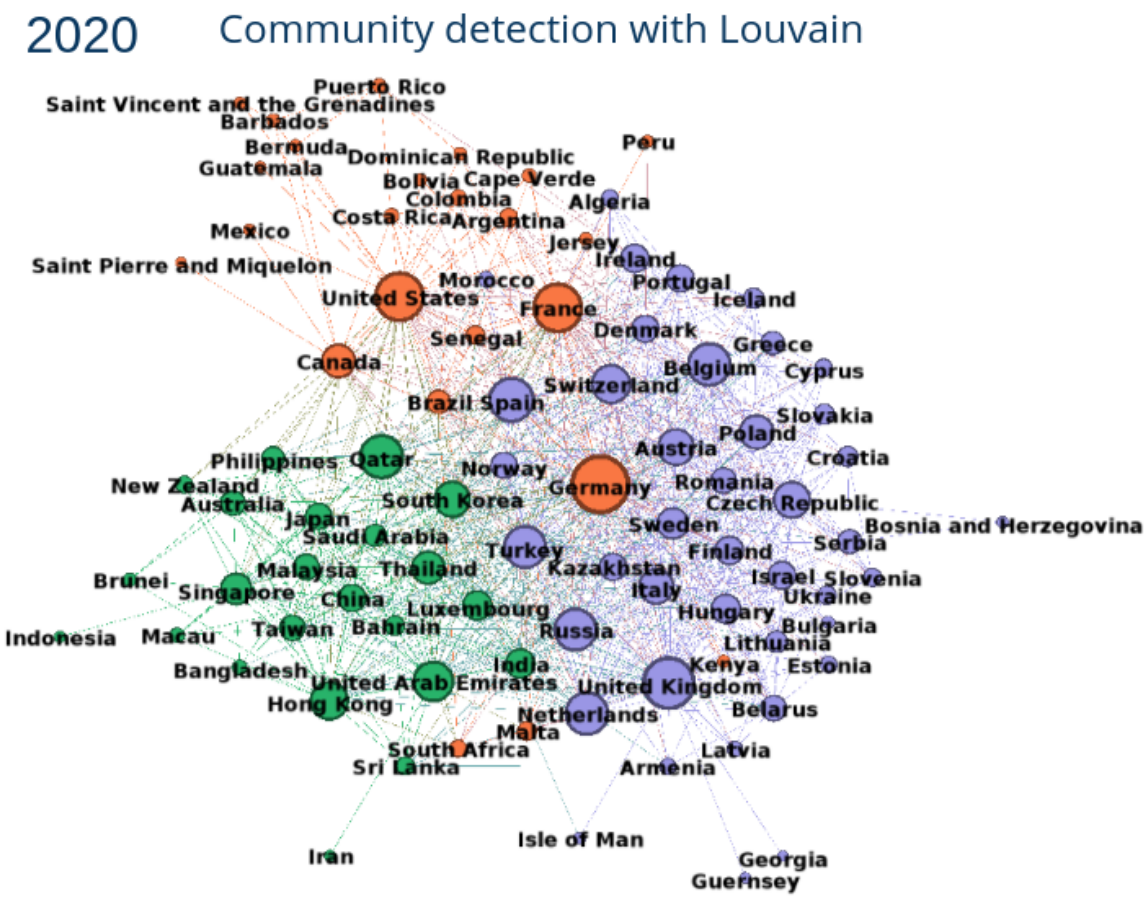


Figure 20b: Community detection Louvain 2020

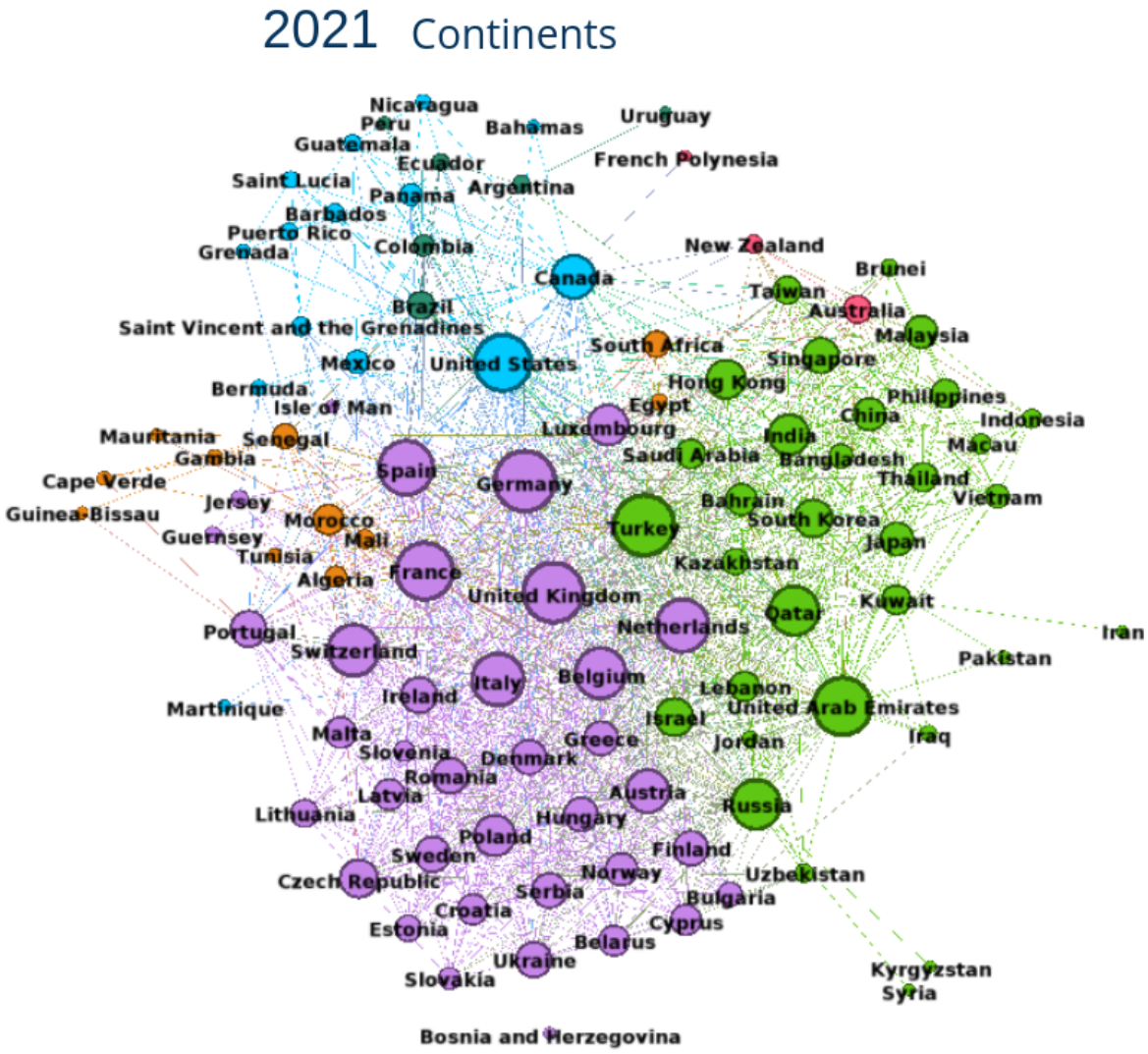


Figure 21a: Continents partition 2021

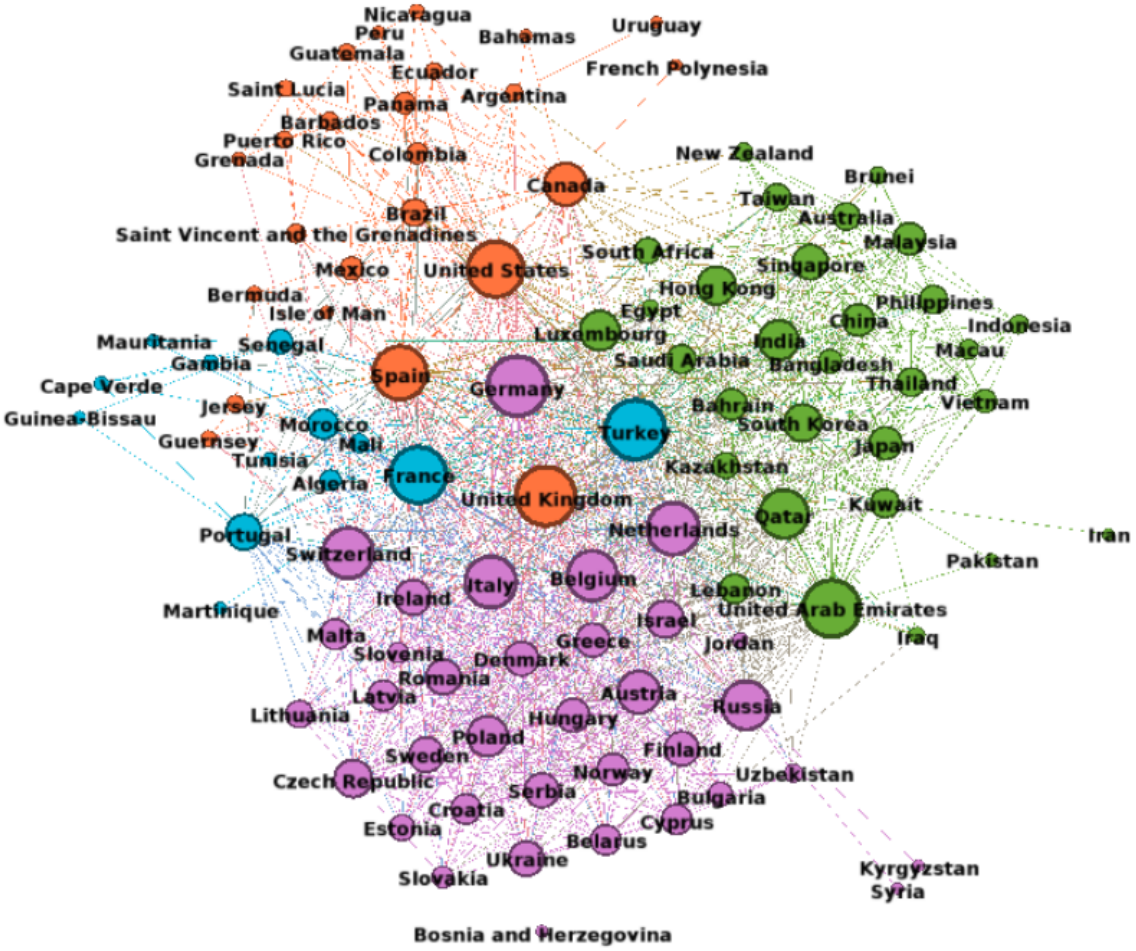


Figure 21b: Community detection Louvain 2021