COVID-19 and the Global Flight Network

Daniel Tang

May 31, 2020

Abstract

Our motivation for this study was to model worldwide flight traffic as a network in tandem with the number of reported COVID-19 cases to map the early spread of the virus internationally. We outline the data sources we used in this study, the steps required to process the data, as well as the main complexities involved when handling the data cleaning process. The degree of connectivity of the worldwide flight network over time was analysed in order to provide insight into the transmission of COVID-19 globally. We observed that regions with a high degree of inbound connectivity are the most susceptible to the initial spread of virus transmission which can put their populations at severe risk.

Contents

1 Data Overview

In order to produce a network model of worldwide flights to integrate with data on the spread of COVID-19, we first needed to have information that detailed each flight between airports over the past six months. This would act as a proxy to measure the relative number of people travelling between two locations at any given time, as actual passenger data was unattainable at both a reasonable price and at a level of detail we required. Recent flight data at the granularity of individual flight counts were also required as an analysis of historical flight data alone would be limited in its usability to model the evolution of the network as confirmed COVID-19 cases increased worldwide.

1.1 Data Sources

We began by downloading the COVID-19 case data from the Johns Hopkins University COVID-19 data repository, which provided a time series data set of confirmed cases, recovered patients, and deaths due to COVID-19 worldwide. Location data was generally specified at the level of each country - with selected regions (such as the United States of America, China, Australia, and some others) offering case numbers at the more detailed granularity of states, provinces, and overseas territories.

Secondly, we had to find a data set that detailed daily flight data between countries over the past six months. We sourced this from the subscription-based data API provided by Aviation Stack [1], which had data on each flight offered by airlines globally. Downloading the daily flight data since December required us to handle a total of approximately $200,000$ flights (each flight a row in the data set) for each day, which meant we had to batch process the download process to deal with memory constraints. Notably, the number of these rows reduced considerably from $200,000$ in December to as low as $50,000$ in May - which is primarily due to shut down of many airline services, but also may be due to the fact that disruptions in the airline industry may have also affected data recording quality and practices.

Geographic information was downloaded from a number of open databases, which were then used to ensure that the merging process was as accurate as possible as we needed to aggregate information by location between data sets. We outline the main data sources used to produce our final analysis below:

- **Historical Flight Data:** Aviation Stack [1];
- **Information on Airports and Airlines:** Aviation Stack [1], Aviation Edge [2], International Civil Aviation Organisation [3];
- **COVID-19 Confirmed Cases:** Johns Hopkins University COVID-19 Data Repository [4];

- **Geographical Identification Data to Merge Datasets:** GeoNames [5]; and
- **Geospatial Vector Data (shapefiles) for Map Visualisations:** Natural Earth Data [6], GADM [7].

1.2 Data Processing

After sourcing our data, extensive data processing was required to merge the data together. In particular, it became apparent that having such a large data set of individual flight records from airports around the world comes at the cost data cleanliness - many rows had either duplicate or missing data, the same airline or airport may have different names or codes according to different sources, and countries and states were sometimes spelt or grouped differently depending on the source. When we did not have a clear solution to handle invalid values, we decided to remove these rows from the data set - however we believe we did not remove too many records so as to preserve the overall structure and dynamics present in our flight data.

1.2.1 Handling Datetimes

We also had to handle datetimes and timezones that are inherent to time series data. Sometimes the records of flight durations did not make sense given the physical distance travelled. Timezones had to be converted manually as they were incorrectly labelled as UTC time from the source API. As an example, when timezones are not handled properly in flight data, a flight from Beijing to Los Angeles would suggest the plane is travelling back in time.

1.2.2 Unique Transits

The granularity for our source data was one row per flight offered by an airline. This meant that when multiple airlines offered flights to customers shared on the same plane, there would be a corresponding number of extra rows in the data set. If a number of companies share a single plane, then using airline flights as the metric would skew our data incorrectly without having actual passenger count data at our disposal. We therefore had to reverse engineer the number of unique plane transits rather than airline flights using data that specified where airlines shared flight numbers.

1.2.3 Computational Efficiencies

When processing the raw data, we also had to ensure that this was done in an efficient manner. Programming simple for loops over each row was not viable, and we tried to use vectorised programming operations as much as possible to increase computational efficiency. We also produced a number of animations for this research and optimised our code where applicable to reduce processing time when building the animation for the full world map.

2 Flight Network Model

We propose a model for the worldwide flight network as a directed weighted graph where nodes in the network represent airports at a specific location, and edges represent directional travel between a specified departure node and an arrival node. Each flight was recorded at some specified datetime, and we aggregated data into daily bins to calculate our edge weights. The weight of each edge in our network corresponds to the number of flights along that edge on a specified date. An ordered list of these networks were stored so that the number of cases at each node location as well as the number of flights between each node could be analysed as a time series.

2.1 Degree of Connectivity

The degree of connectivity for a directed weight graph can be calculated in terms of the number of inbound or outbound edges per node, known as the 'in degree' and 'out degree' respectively. In our case, this calculation was weighted by the number of transits along each edge daily. Therefore, given a specified date, the degree of connectivity for a node corresponds to the frequency of flights either departing or leaving an airport each day in our model.

Below we show subplots that have the mean in degree and out degree over time across all airports in our data set. As the degree of a node corresponds to the frequency of flights at that node, this simply reveals the extent to which flight traffic has dropped since the beginning of the year.

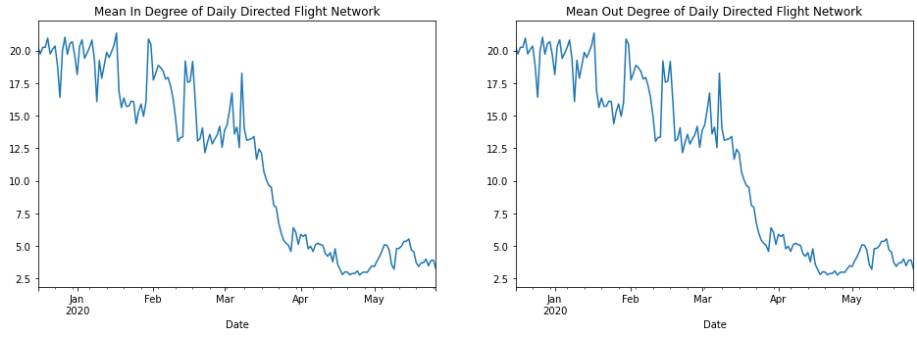

Figure 1: In degree and out degree over time averaged across all airports globally.

Above we can see that generally the inbound and outbound flights have decreased somewhat equivalently. As the focus of our study is to later look at inbound flight frequency and the risk of COVID-19 spread, we generally focus

on the in degree moving forward.

Below we show the mean in degree of network nodes every week grouped by continent. The weekly aggregation is helpful to average out effects of dropping rows during data cleaning and to smooth out the curve that appears quite noisy at the daily scale. In addition to the mean in degree, we also show the percentage change by continent since the 16th of December 2019.

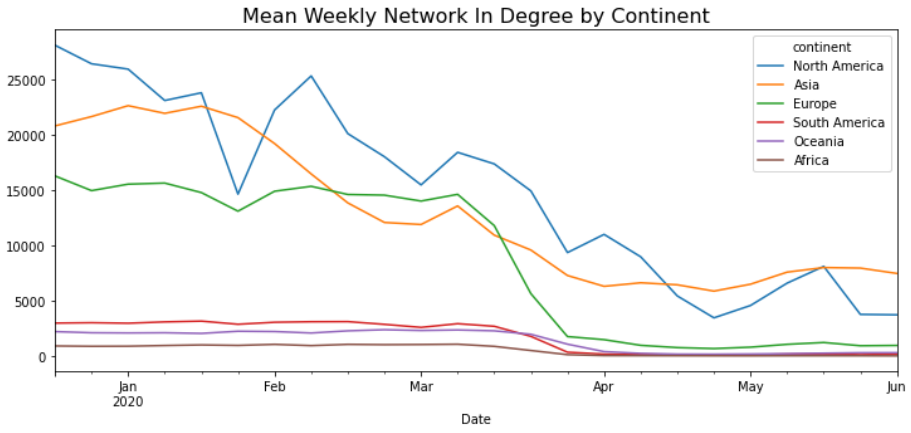

Figure 2: Weekly in degree by continent.

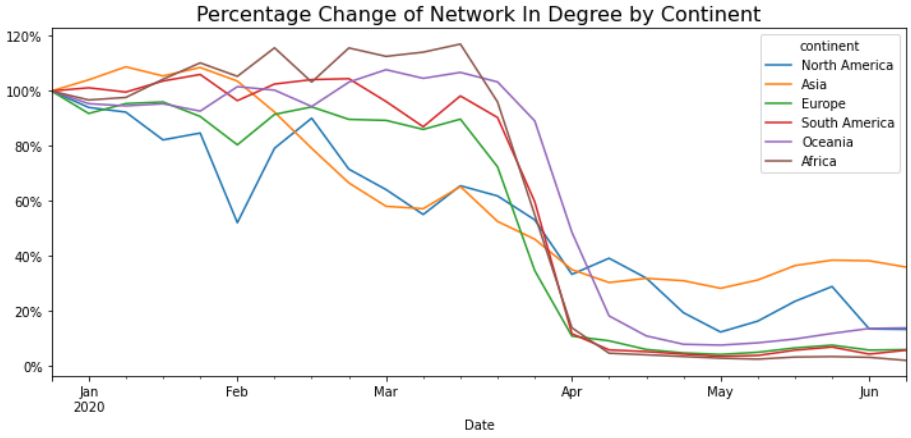

Figure 3: Percentage change of in degree by continent.

From the plots above we can see that North America and Asia appear to the primary hubs of flight network traffic, followed closely by Europe. Secondly, a

large drop in flight traffic occurs in late March onward, as airline services and airports reduce activity in response to the global pandemic.

2.2 Degree Distributions

The distributions of degree for our nodes shown above may reveal some characteristics about the structure of our network. In particular, if the degree distribution has a heavy-tailed distribution, it can suggest that there is a power law relationship between the most connected nodes in the network and their probability of occurrence. It is easier to determine if our degree distribution has a power law relationship by plotting them on a log-log axis and seeing if under this transformation the data falls on a straight line [8]. The figures below show that this relationship holds both at the start and at the end of our time series, with some non-trivial variance at the higher end of our degree distribution.

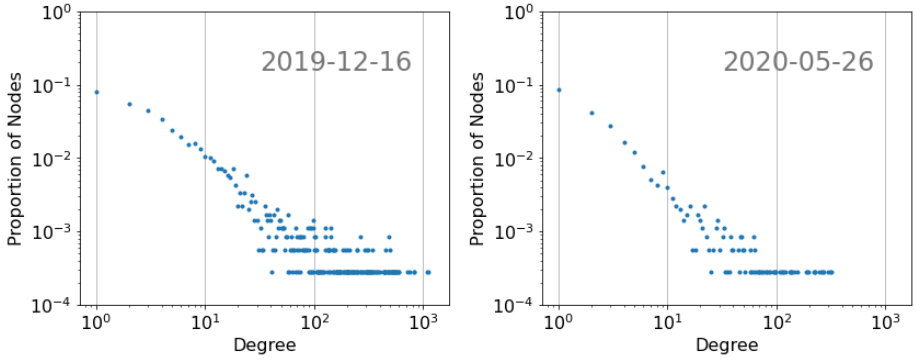

Figure 4: Network degree distribution on log-log scale on selected dates.

In a paper published by Albert-Laszlo Barabasi and Reka Albert in 1999 [9], a network model is proposed that generates graphs with a heavy-tailed distribution for the degree of nodes in a network. They posit that phenomena that have patterns of growth and preferential attachment often lead to networks with power law relationships. Networks with this property are often called scale-free networks [8].

We posit that the global flight network is one of such scale-free networks. Hubs of transit activity have emerged after many years of growth in the airline industry, alongside market forces for preferential attachment to established airlines and airports. Preferential attachment does not have to come about by consumer preference - we can imagine airports functioning as centralised transit hubs for local regions being desirable from the perspective of streamlining transport operations, ease of passenger access, as well as optimising for financial

and business constraints.

When we consider the degree distribution subplots over time alongside the large drop in flight traffic as displayed before - we can see that this heavy-tailed distribution in the network remains robust even as the number of flights (degree) has decreased by approximately 80% worldwide.

3 COVID-19 Spread

The COVID-19 reported cases data was sourced from the Johns Hopkins University GitHub page [4], which details the number of reported cases since the 22nd of January 2020. The number of reported cases around the world accelerated rapidly in March and April after the situation in Wuhan, Hubei Province in China was not entirely contained. Below we show some plots on both a linear and a log scale to emphasise the rate of increase in confirmed cases globally.

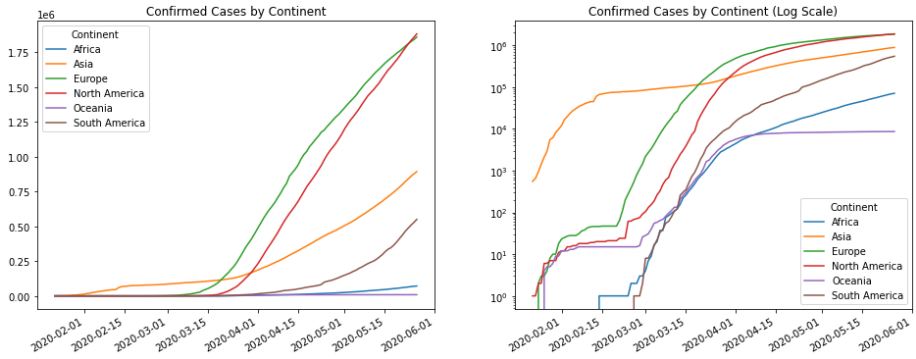

Figure 5: Confirmed cases over time grouped by continent.

3.1 Regions at Risk of Virus Transmission

Our initial analysis of our flight network model revealed that North America, Asia and Europe were the continents with highest count of flight traffic in Figure 2. We were able to narrow our analysis to the country level to see which countries had the highest inbound connectivity and we chose the subset of these countries sorted by in degree for analysis. As the United States of America was evidently the country node with the highest degree, we were able to split them up on a state level - states such as California, New York, and Texas all had a degree comparable with countries internationally. We selected fifty of the top international countries and US states by in degree to model whether or not these locations were the most susceptible to a high number of cases and early

transmission of the virus. This snapshot was taken at the beginning of our time series to minimise the effect that disruption to the airline industry may have had in the order of these rankings.

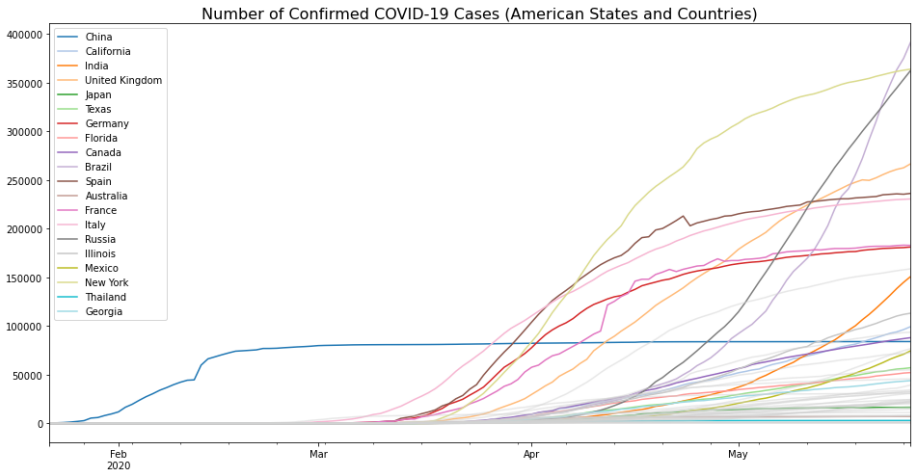

Figure 6: Confirmed cases over time for regions with high in degree.

The plot above shows the confirmed cases for these selected regions, highlighting the top twenty out of the focus group to emphasise those regions most at risk to virus spread. We can clearly see that these highlighted regions are generally reporting a higher number of confirmed cases whose growth has often started earlier than other countries.

A visualisation of the world map where our focus regions (with higher risk due to higher connectivity) were highlighted also shows this trend of virus transmission spreading to these countries first. The number of cases in Europe and North America are the highest, with regions in Africa and Oceania having a comparatively lower count. It is notable that in North America the virus first spread to coastal regions that have higher population count and more frequent flight traffic. A number of countries in Europe such as Germany, Italy, France, Spain, and the United Kingdom were also at high risk. Brazil was also one of the worst hit countries in South America, with both the highest degree of connectivity and population in the continent. The details of our selected regions and their cases numbers are provided in our code notebook, and snapshots of our animation are displayed in the appendix section.

3.2 Discussion

Modelling global flight traffic as a time series network has revealed valuable insights about susceptibility to risks that may traverse the topology of the worldwide airline network. In particular we note that the same heavy-tailed distribution that aptly models the degree of connectivity in our flight network often well describes the risk and unpredictability of large scale destructive events. Both the rate of transmission and the associated mortality of COVID-19 has been severe in most cases, and the speed at which cases were reported across international waters suggests that the global flight network would be the primary candidate as the medium of travel for virus transmission in modern day. As urban population centres around the world grow and build airports to facilitate international transit activity, we must be wary that this too increases their degree of connectivity and therefore vulnerability to global destructive phenomena that may emerge from the larger world around them.

3.3 Further Research and Related Work

Other papers have also studied the worldwide airline network to find heavy tailed distributions in connectivity, small world properties, and network clusters that function as either central transport hubs, far-reaching periphery nodes, and bridge clusters that exist between them [10], [11]. An analysis of virus transmission that incorporates the ideas thoroughly explored in these papers would reveal much about the structure of the worldwide airline network as a medium for the transmission of more than just the planes and their passengers onboard.

4 Appendix

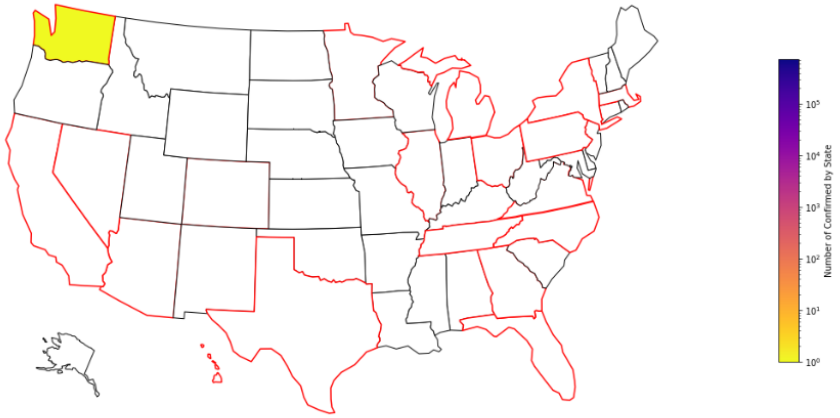

Figure 7: COVID-19 cases reported in USA with high in degree states outlined in red (January 22 2020).

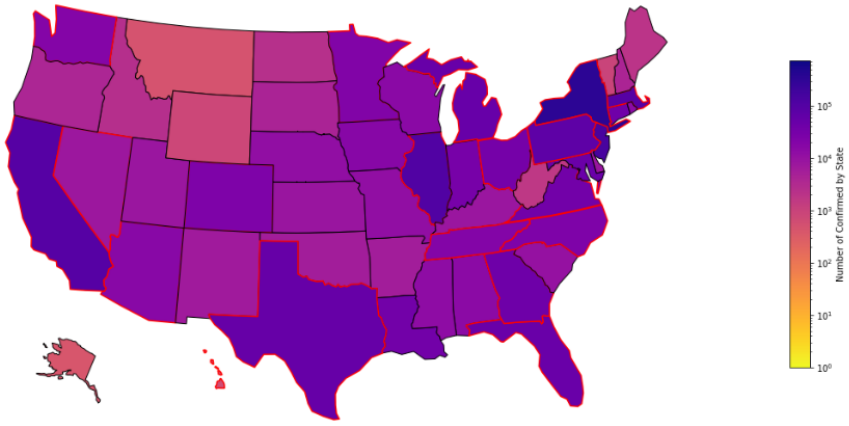

Figure 8: COVID-19 cases reported in USA with high in degree states outlined in red (May 25 2020).

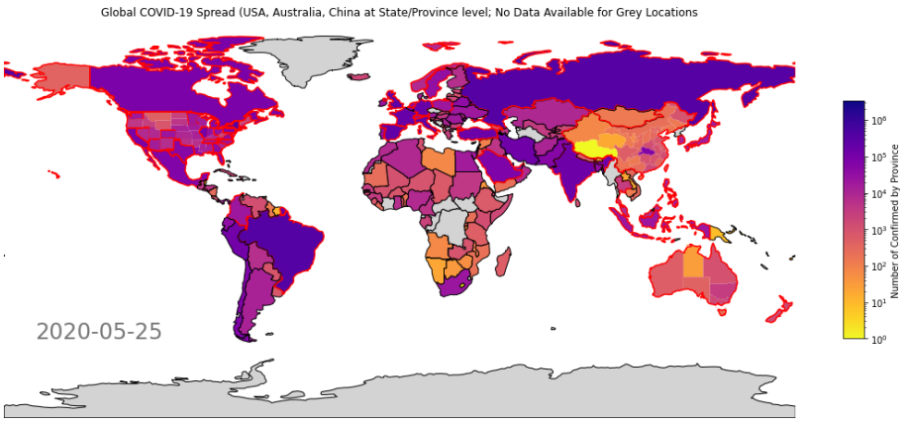

Figure 9: COVID-19 cases reported globally with high in degree regions outlined in red (May 25 2020).

References

[1] Aviation stack. `https://aviationstack.com/`.

[2] Aviation edge. `https://aviation-edge.com/`.

[3] International civil aviation organization. `https://icao.int`.

[4] John hopkins university covid-19 data repository. `https://github.com/CSSEGISandData/COVID-19`.

[5] Geonames geographical database. `geonames.org`.

[6] Natural earth data. `http://naturalearthdata.com/`.

[7] Gadm. `gadm.org`.

[8] A. B. Downey. Think complexity 2nd edition. O'Reilly, 2018.

[9] Albert R. Barabasi, A. Emergence of scaling in random networks.

[10] Araújo N. A. M. Verma, T. and H. J Herrmann. Revealing the structure of the world airline network. `https://www.nature.com/articles/srep05638`, 2015.

[11] Mossa S. Turtschi A. Guimerà, R. and Amaral L. A. N. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. `https://www.pnas.org/content/102/22/7794`, 2005.