Graph Analysis of Air Transport Network

Assignment – Data Analytics and Visualisation

Deepak Kovaichelvan – S336570

Date: 4th January 2022

Information categorisation: Open

Table of Contents

1. Airline Network Visualisation

The objective of the assignment is to perform network data analysis on the domestic airline networks for 4 countries of interest – USA, China, UK and Australia.

The supporting dataset for the assignment included the flight data for various countries, spanning both domestic and international routes. The flight data was imported into Python in the form shown in Table 1.

Each row of the table is an edge of the network, as it includes information of the source and target airports between which the routes operate, along with TimeSeries data and weights, which signify the available seat capacity on the flight.

The source and target determine the direction of the flight route. Therefore, each row represents a weighted directed edge of the overall flight network. There are some source and target airports between which multiple flights operate. To accommodate this, the network was treated as a multiple directed network, meaning multiple edges can exist between any two nodes.

Table 1: Top 5 rows of the Flights DataFrame

	Source	Source City	Source Country	Target	Target City	Target Country	Weight	TimeSeries
0	FNC	Funchal	Portugal	PXO	Porto Santo	Portugal	9864	2003-07-01
1	PXO	Porto Santo	Portugal	FNC	Funchal	Portugal	9864	2003-07-01
2	AEP	Buenos Aires	Argentina	MVD	Montevideo	Uruguay	1463	2003-07-01
3	MVD	Montevideo	Uruguay	AEP	Buenos Aires	Argentina	1463	2003-07-01
4	AEP	Buenos Aires	Argentina	ROS	Rosario (AR)	Argentina	2261	2003-07-01

Filters were added to the original DataFrame to extract domestic flights for each of the 4 countries and July of 2007 was selected as the month of interest for the purpose of this assignment.

The NetworkX package on Python was used to convert the list of edges contained in the Table 1 to a graph form. The graph only encodes the connections between nodes ignoring their geographic positions, which is particularly relevant in transportation networks as the nodes are located relative to each other in geographic positions.

A supplementary dataset of airport coordinates was used to extract the latitudes and longitudes of each airport. The positions of the nodes and their corresponding edges were projected on to a map using the Cartopy package.

The resulting domestic airline networks of the 4 countries are shown in Figure 1 to Figure 4.

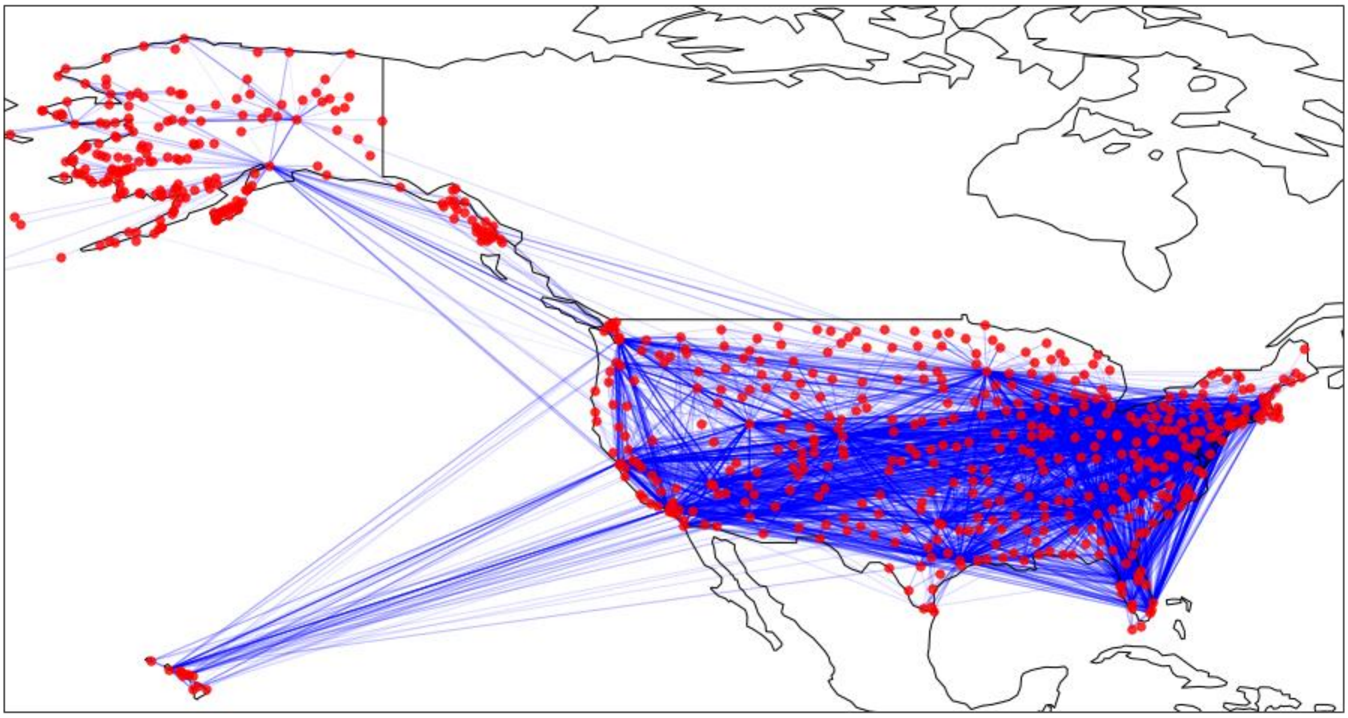


Figure 1: Domestic airline network of the USA

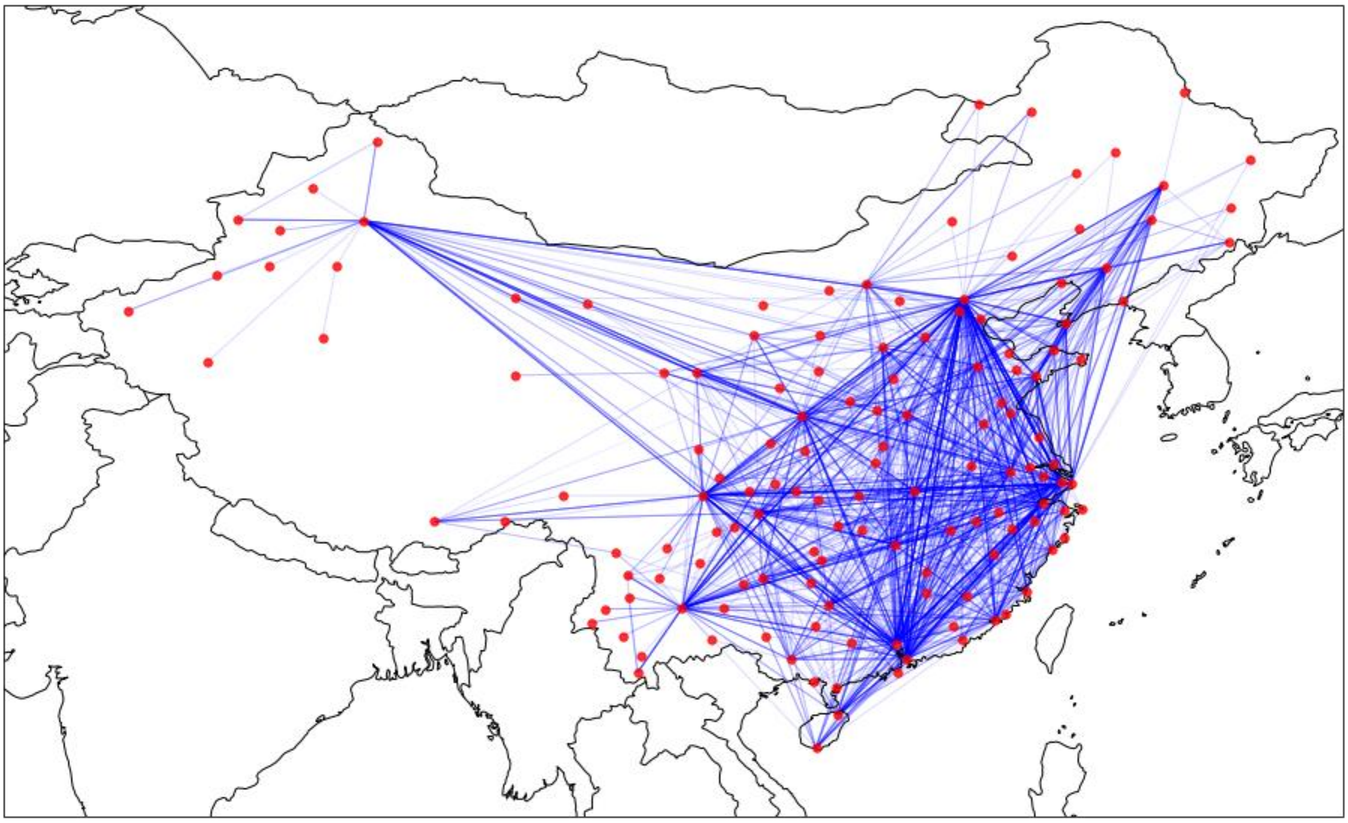


Figure 2: Domestic airline network of China

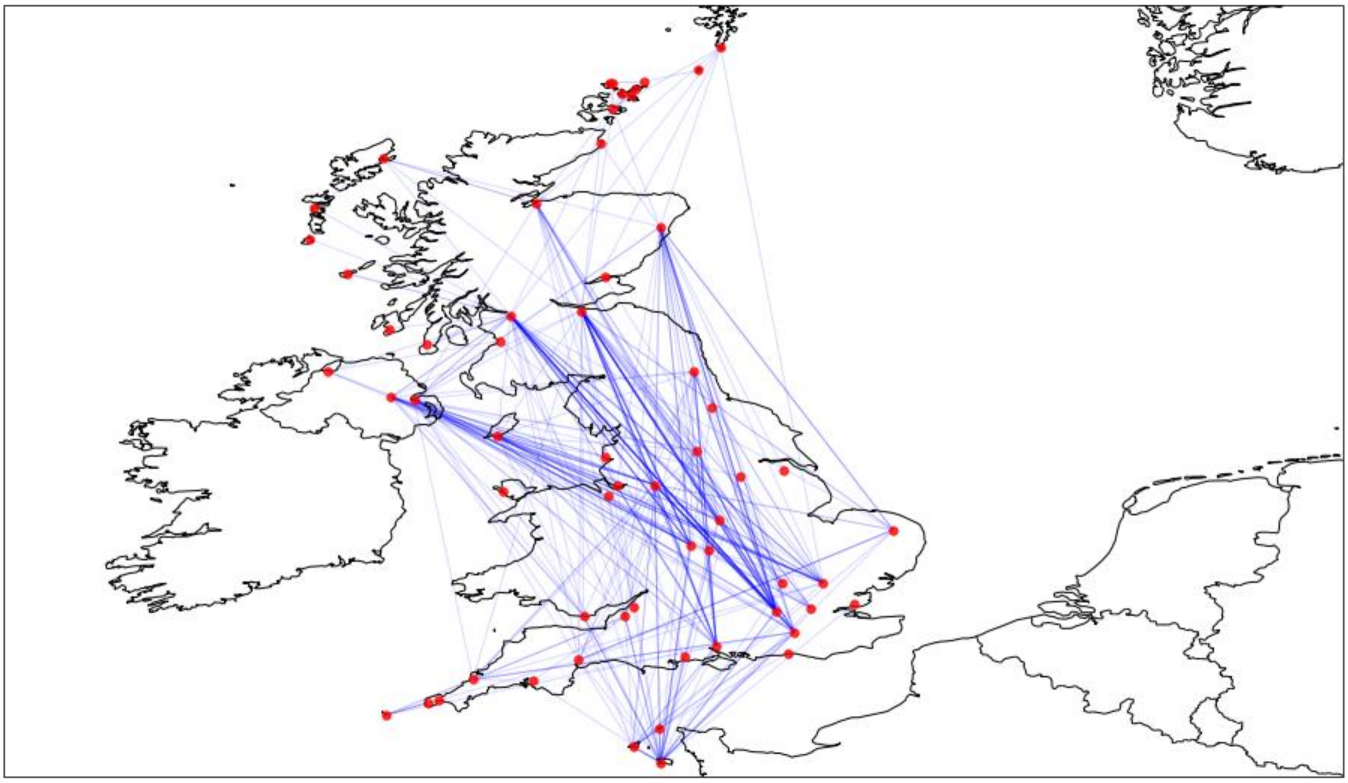


Figure 3: Domestic airline network of the UK

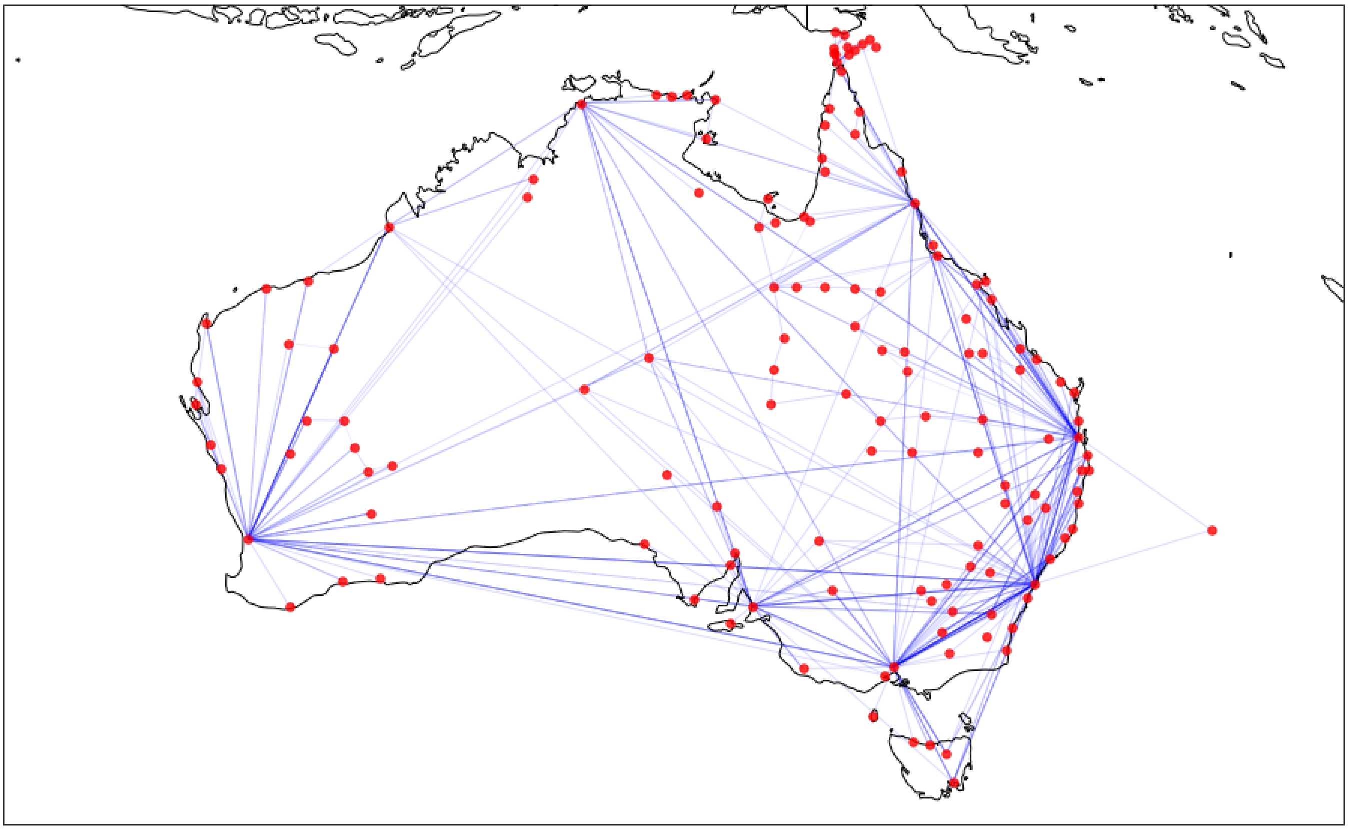


Figure 4: Domestic airline network of Australia

2. Plots

2.1. Degree Distribution

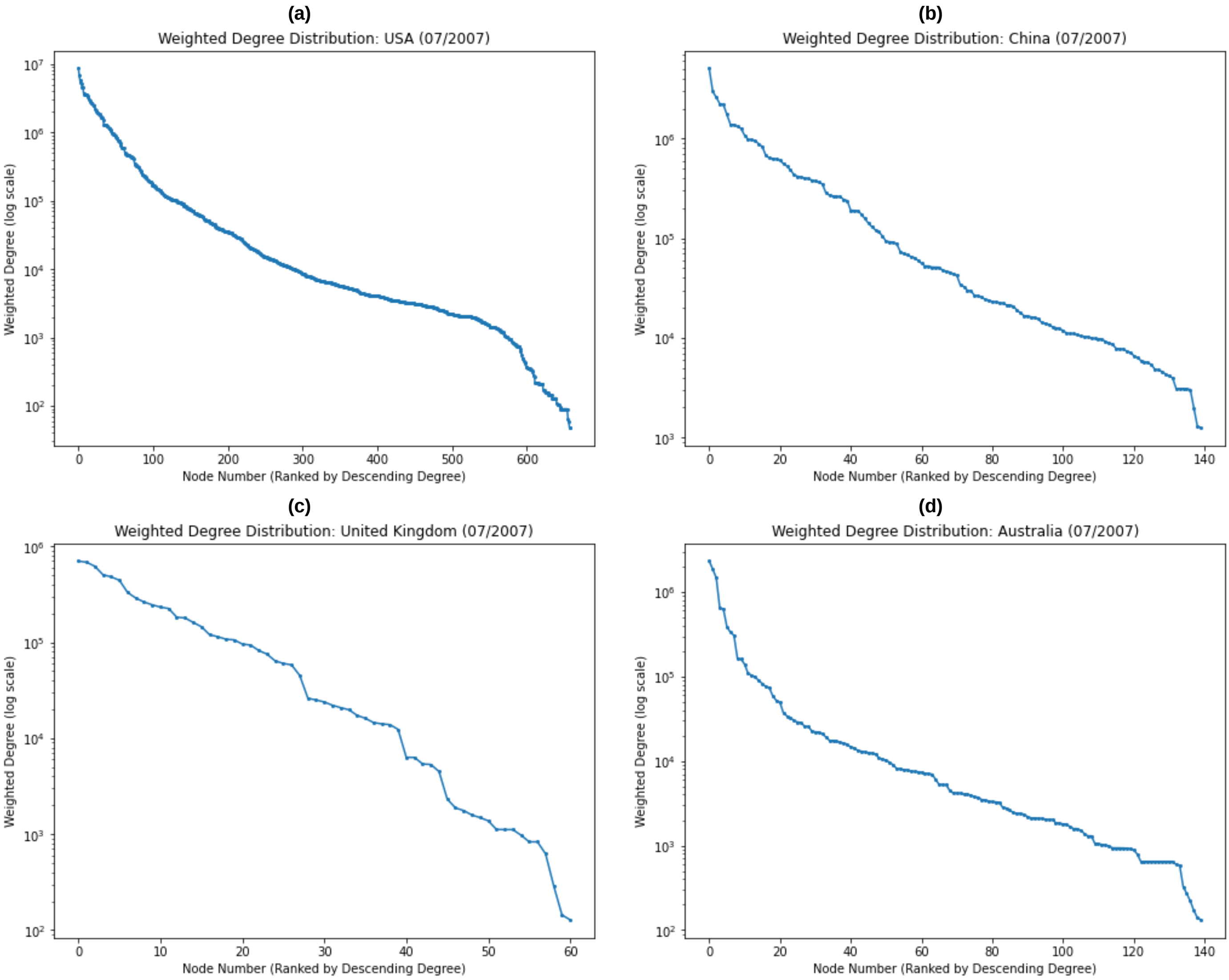


Figure 5: Weighted degree distributions (a) USA, (b) China, (c) UK and (d) Australia

Figure 5 shows the weighted degrees of nodes (in log scale) plotted against their node numbers ranked by descending degree. The weighted degree of a node is the sum of weights of the edges connected to the node.

2.2. Degree vs. Betweenness Distribution

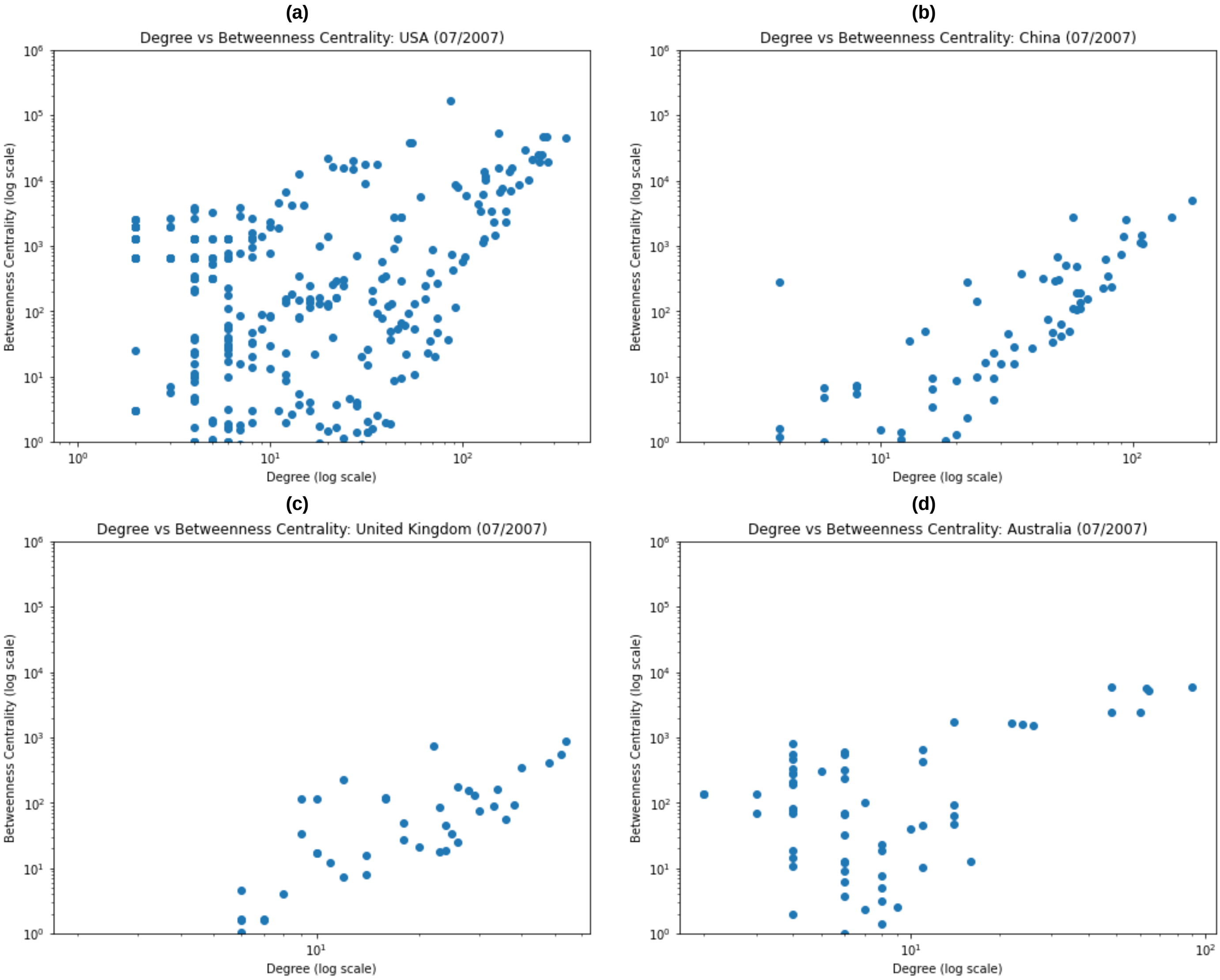


Figure 6: Degree vs betweenness centrality plots (a) USA, (b) China, (c) UK and (d) Australia

Figure 6 shows the degrees of nodes plotted against their betweenness centrality scores, both shown in log scales. Degree of a node is the number of edges connected to the node. Betweenness centrality is a measure of the extent to which a node is in the path between other nodes [1].

2.3. Assortativity (Degree Correlation)

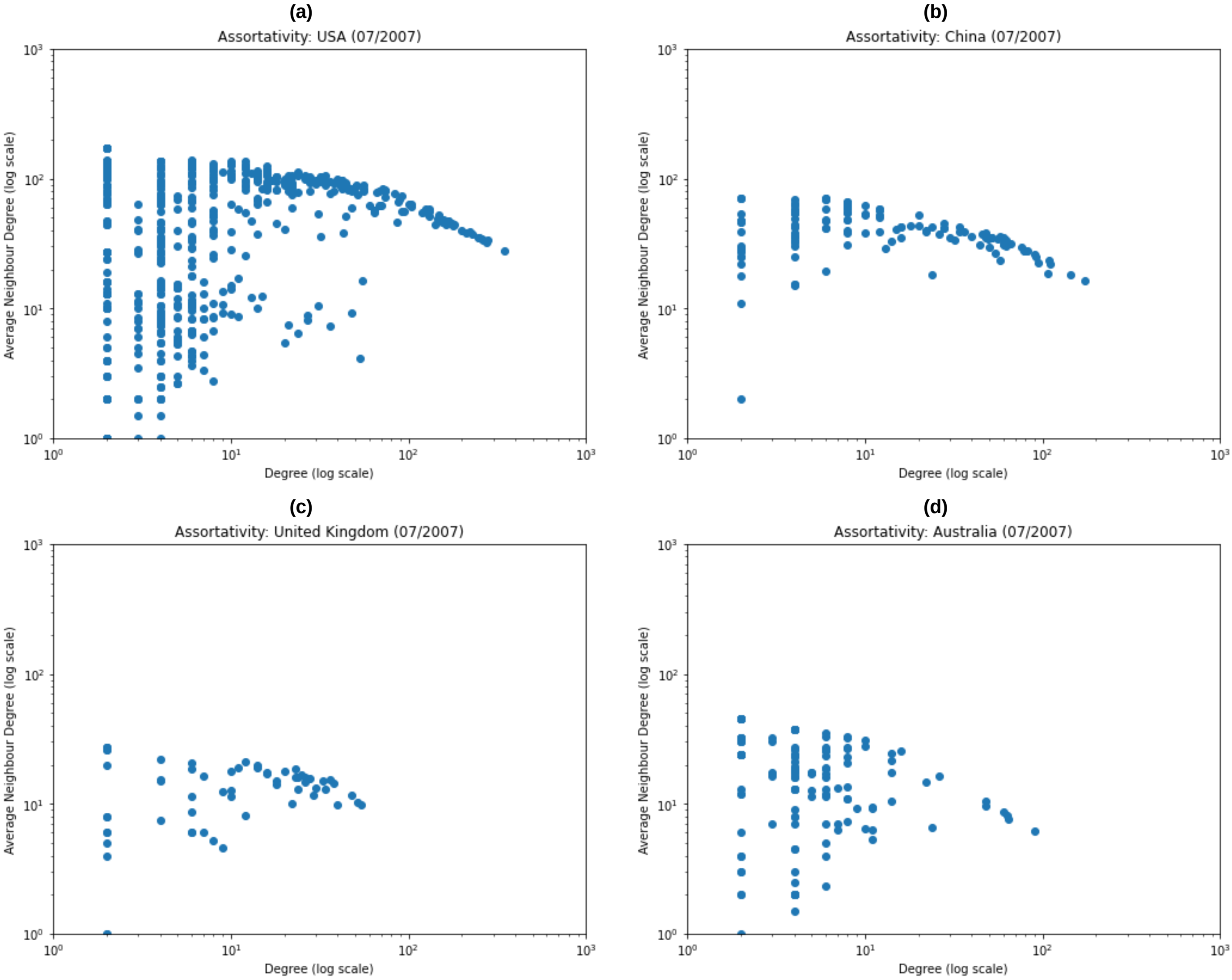


Figure 7: Average neighbour degree assortativity (a) USA, (b) China, (c) UK and (d) Australia

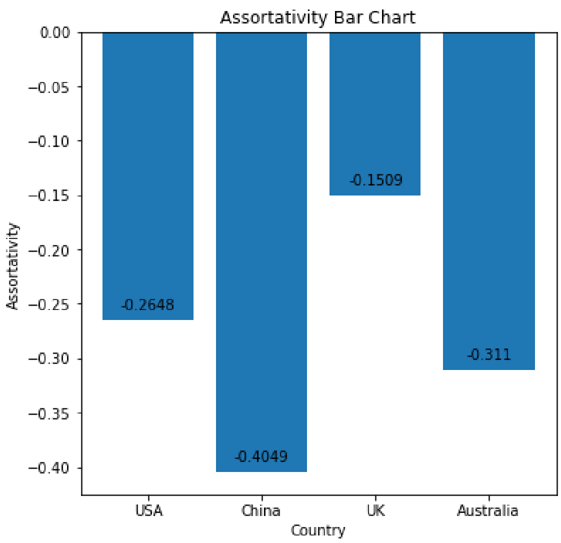
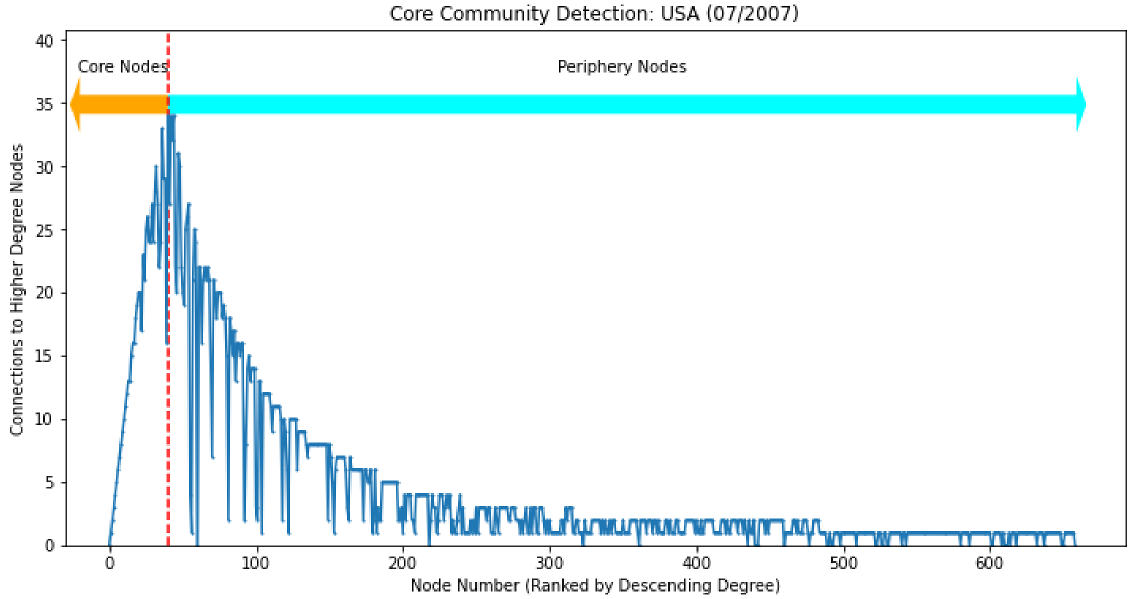


Figure 8: Comparison of assortativity coefficients by country

Figure 7 shows the degrees of nodes plotted against the average weighted degree of their neighbours. Neighbours of a node are nodes that have an edge shared with the node of interest. Figure 8 summarises the assortativity coefficients of the countries of interest.

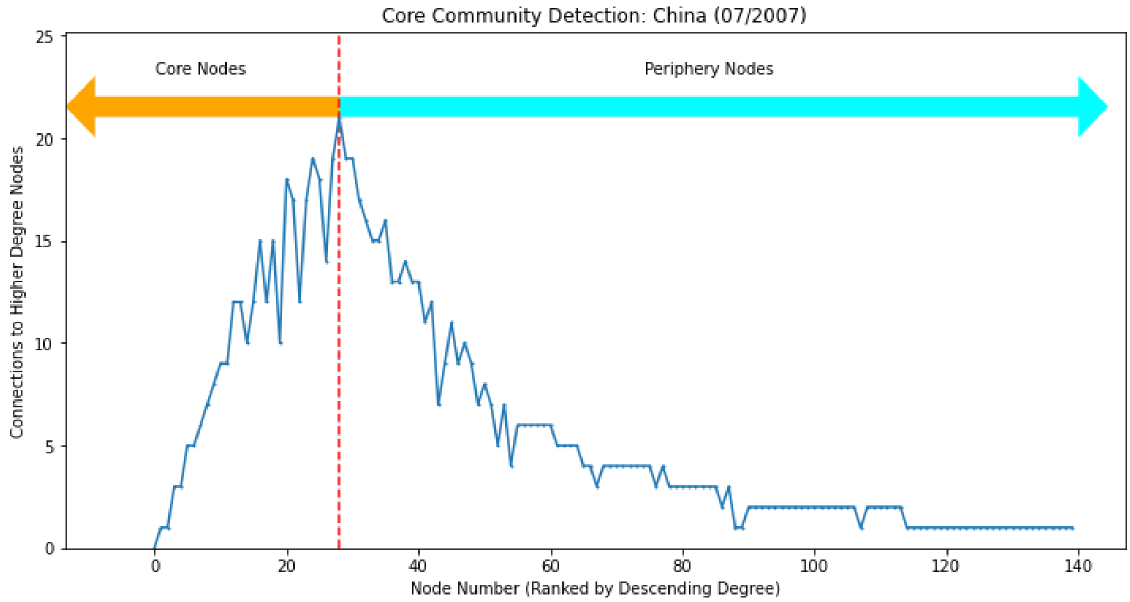
2.4. Core Community Size

(a)



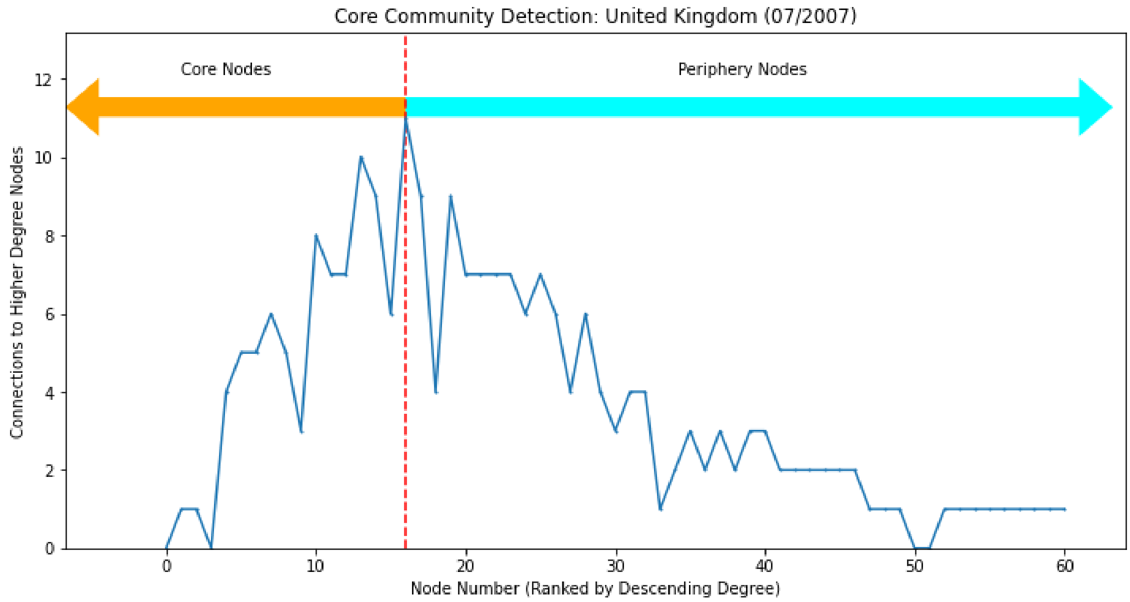
Core Size: 40

(b)



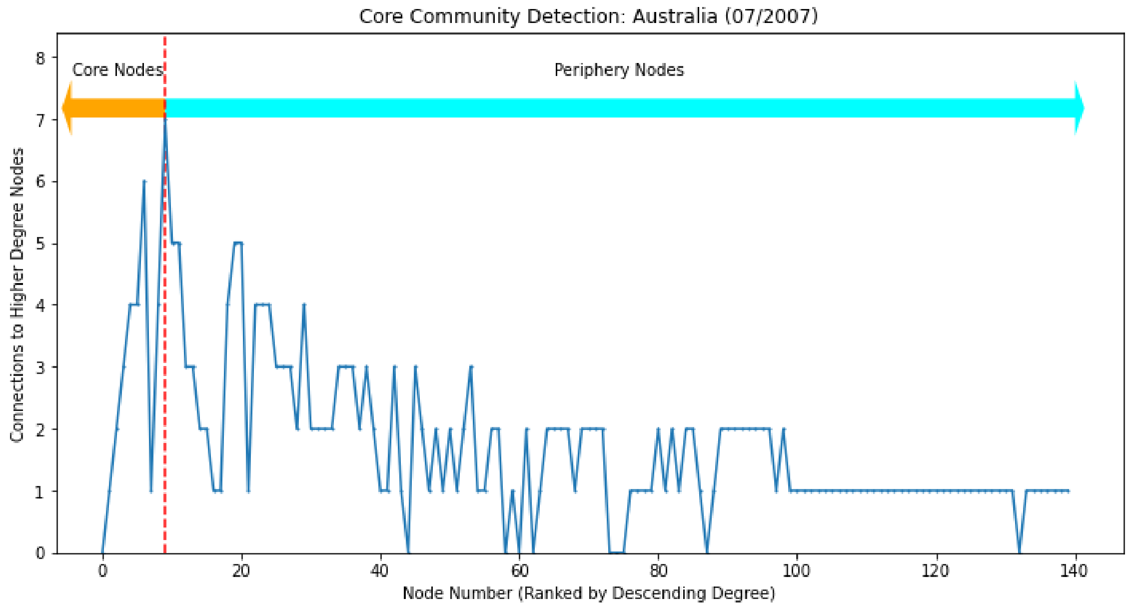
Core Size: 28

(c)



Core Size: 16

(d)



Core Size: 9

Figure 9: Core and periphery separated at the maximum number of connections to higher degree nodes for (a) USA, (b) China, (c) UK and (d) Australia

Figure 9 illustrates the separation of the core and periphery nodes at the node number with the maximum number of connections to higher degree nodes.

3. Analysis

3.1. Degree Distribution

Figure 5 shows that the weighted degree distribution follows the power-law distribution. This means that the fraction of nodes with degree k varies as a power of degree k [1]. While a large proportion of the variance can be explained by the exponential distribution, there exist King and Pauper effects at the ends of the node ranks. This means that the first few ranks have an order of magnitude higher weighted degree (King effect) and the last few ranks have an order of magnitude lower weighted degree (Pauper effect) [2]. The King airports are those that play the role of hubs in the hub-and-spoke (HS) system. This effect is most visible in the USA and Australia.

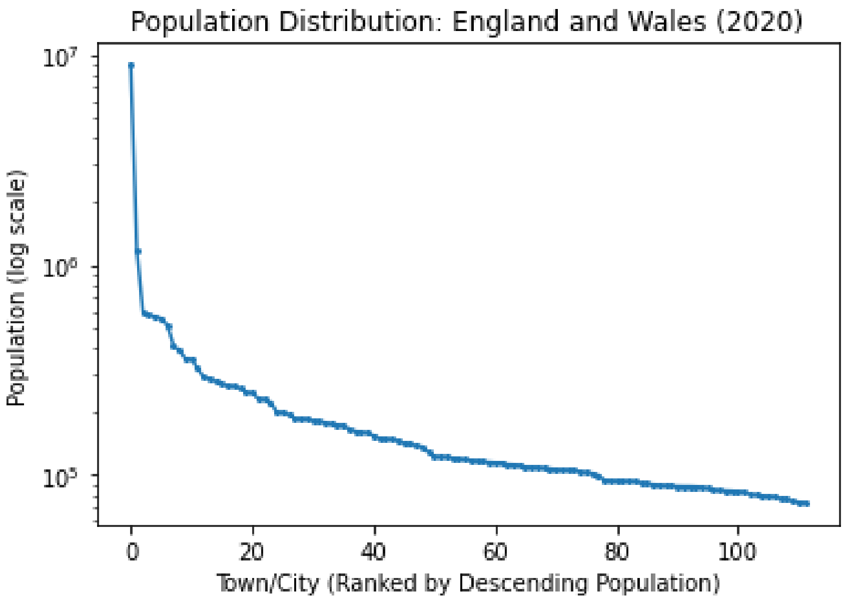


Figure 10: Population Distribution of England and Wales (majority of the UK's population)

The degree distribution was compared to the population distribution for England and Wales (regions which constitute the majority of the UK's population) using data from the Office for National Statistics (ONS) [3]. Figure 10 shows that the population distribution follows a similar exponential distribution to the degree distribution seen in Figure 5. Notably, the Pauper effect is missing as the ONS data only covers major towns and cities; including the smaller towns and villages is expected to have made the Pauper effect visible.

3.2. Degree versus Betweenness

The most interesting observation from Figure 6 is the convergence/sharpening of the betweenness values for the higher degree nodes, compared to the relative scatteredness of the lower degree nodes. The higher degree nodes always have high levels of betweenness.

This indicates that the higher degree nodes also fall in the shortest path between other nodes. This is because the higher degree nodes play the role of transiting hub airports, which are connected to several smaller regional airports. These higher degree nodes therefore play a critical role in connecting the network, as do a small number of the lower degree nodes which also have high betweenness values.

3.3. Assortativity

Assortativity is a key indicator of the extent to which a network's nodes have the tendency to attach to nodes that are similar to themselves. In this case, the assortativity has been studied with respect to node degree. Figure 7 shows that the assortativity (by average neighbour degree) at lower degree nodes is quite scattered (neutral correlation) and steadily becomes disassortative at the higher degree nodes due to the negative slope.

This is because there are some regional airports that connect to other regional airports, while some other regional airports mostly connect to the larger hubs. In contrast, there are far more edges from hubs to regional airports than to other hubs, resulting in disassortativity at the tail of the graph.

Figure 8 demonstrates from the negative assortativity coefficients that all the 4 networks are disassortative, which suggests that the nodes attach to other nodes that are dissimilar in terms of degrees. This is mainly due to the disassortativity seen at the higher degree nodes in Figure 7. Although the overall coefficients point to disassortative networks, they are only partly disassortative.

China has the highest disassortativity and the UK has the least disassortativity.

3.4. Core-Periphery Structure

The most commonly known network structures are assortative and disassortative communities. These can be described in terms of probability of edges between and within the nodes of the network. For two groups - 1 and 2, in an assortative structure, the probabilities are given by $p_{11} > p_{12} < p_{22}$. While in a disassortative structure, it is given by $p_{11} > p_{12} > p_{22}$ [1].

The third type of structure is a core-periphery structure, the probabilities are given by $p_{11} < p_{12} < p_{22}$. In practical terms, this is a network which has densely connected nodes forming a core, surrounded by periphery nodes, which are moderately connected to the core and sparsely connected to one another [1].

Since it is known that the networks are neither fully assortative nor disassortative, a core-periphery detection algorithm was used. The algorithm arranges the nodes by degree (k) and counts the number of connections to higher degree nodes (k^+) for each of the nodes. The peak of the distribution signifies the core size, as this is the point up to which the connections to higher degree nodes increases [2].

In Figure 9, the boundaries between the core and periphery nodes have been drawn for the countries of interest to show the sizes of the core communities. It can be seen that the USA has the highest number of airports and the largest core size; however, the core forms the smallest proportion of the overall network, along with Australia. This illustrates the high dependency of the networks on a small proportion of core airports.

3.5. Performance Considerations

The performance of networks in a core-community structure is tuned for economic efficiency, as it can be seen that the design intent is to minimise the number of routes and maximise the economies of scale at the large hub airports. This can however lead to lower robustness to failures of the whole network, due to the high dependency on a small number of core nodes/hub airports. Failure of these core nodes is likely to result in major disruption to the whole network.

3.6. Fuel Price Impact on Network Model and Future Aircrafts

The networks models are highly influenced by the economics of the airline operations. A key function to understand the economics of networks is the probability of forming a link between two nodes, i and j in a random geometric graph. This function is given by the below equation [2].

$$Q_{i,j} = K{d_{i,j}}^{-\propto}$$

Where, K is a normalising factor (constant), d is the distance between the nodes and α is the fuel cost coefficient.

It can be seen from the above equation that as the distance and fuel cost coefficient increase, the distance cost increases, reducing the probability of forming an edge. This is because the distance penalty dwarfs the benefit of the economies of scale of the hub-spoke system. This would mean that the networks would move from a hub-spoke (HS) model to a point-to-point (PP) model.

Since in the PP model, flights operate between smaller regional airports, it would not make economic sense to operate large aircrafts due to the difficulty in filling all the seats when operating from small regional airports. Therefore, the preference would shift to small and medium size aircrafts [2].

4. References

[1] M. Newman, Networks: Second Edition, Oxford: Oxford University Press, 2018.

[2] W. Guo, B. Toader, R. Feier, G. Mosquera, F. Ying, S.-W. Oh, M. Price-Williams and A. Krupp, “Global air transport complex network: multi-scale analysis,” *SN Applied Sciences,* 2019.

[3] Nomis, Office for National Statistics, “Population estimates - small area based by single year of age - England and Wales,” [Online]. Available: https://www.nomisweb.co.uk/query/construct/components/stdListComponent.asp?menuopt=12&subcomp=100. [Accessed 02 January 2022].

Appendices

Appendix A – Python Code

The Python code used to generate the graphs and metrics is available to download from the below URL:

https://drive.google.com/file/d/1TPDFFrnK5C7fZgPyGWOs8WK5cs3yVzX7/view?usp=sharing