Uncovering the topological structure of the World Airline Network

Tshepo Yane

This manuscript was compiled on December 2, 2021

The World Airline Network (WAN) plays an integral role in connecting societies across the world. It is therefore important that we understand its structure to improve its efficiency, anticipate future failures and problems and devise solutions. Currently, there are no comprehensive studies that provide frameworks that reveal the topological structure of the WAN. If the day comes whereby the WAN fails, we have no guidelines based on empirical evidence to react accordingly to the event. Luckily, the WAN can be modelled as a complex network with nodes and edges. In this study, We use different centrality metrics to rank the importance of different airports in the network, we uncover the community structure of the WAN and the overall macroscale organizational structure of the WAN. Furthermore, we discover the clustering of different airports is not fully dependent on geographical location.

INTRODUCTION

Mankind has been migrating across the planet since the beginning of times. Over the years, humans developed numerous modes of transportation to allow for easier and faster travel across the globe. In today's world, transport systems are vital components of human society as well as its economy. These transportation systems provide high levels of mobility, which are vital for the cohesion of different markets and quality of life of different populations (1). Furthermore, transportation systems enable socio-economic growth, and facilitate job creation by connecting individuals to job opportunities that may have not previously been within their reach. For example, the air transportation industry is esimated to have generated 32 million jobs worldwide in 2008, of which 5.5 million were direct, and contributed 408 billion USD to the global gross product (2). In addition to the latter fact, 25% of all companies' sales appear to be dependent on-air transport, while 70% of businesses report that serving a bigger market is a key benefit of using air services (3). From these examples, it can be observed how the Word Airline Network (WAN) is a paradigmatic example of an essential transportation network with a global extension. Before the Covid-19 pandemic, the estimated number of scheduled passengers boarded by the global airline industry corresponding to about 5 billion people in 2019 (4). With the increasing demand for air transportation, this creates an enormous pressure on the WAN infrastructure. In order to understand how we may be able to ease this pressure and foster sustainable development the current WAN infrastructure, an in-depth analysis of this network needs to be performed.

Graph-theoretical studies, implementing network-based analytical tools, can report metrics pertaining to individual nodes (nodal measures), allow one to investigate mesoscale properties or to focus on marcoscale properties such as the network's global efficiency or small worldness. Graph-theoretical tools can be readliy applied to the analysis of the WAN and better understand its structure and dynamics (6). In addition, Considerable research has been conducted on the definition of models and algorithms that enable one to solve problems of optimal network design (7). However, a worldwide, "system" level analysis of the structure of the air transportation network is still lacking. In this study, we aim to fill this gap.

The current organizational features of the WAN are mostly determined by the concurrent actions of airline companies, either private or national, that try, in principle, to maximize their profit. In addition, the architecture of the WAN is also the outcome of numerous geographical, political, and economic factors. Figure 1 shows the spatial arrangement of the different airports that make up the WAN. Currently, there are about 40 000 aiports in the world (8). Nevertheless, the intricate nature of the WAN has been mostly analyzed topologically at the mesoscale (6). Although degree distribution and other centrality correlation are commonly used to characterize a network's organizational properties, researchers found that the abovementioned indicators hardly reflect the influence and significance of the individual nodes or modules (9). Numerous studies have analyzed the topology of WAN via reporting mesoscale properties in isolation. Very few of these studies, however, have conducted a multi-scale analysis of the WAN, by simultaneously investigating measures spanning from mesooscale to marcoscale. One study performed by Guimerà et al. identified communities in the WAN and demonstrated the multi-community structure of this worldwide network (10). Their analysis showed that the community structure cannot be explained exclusively based on topographical restraints, but some geopolitical concerns should also be taken into account. Numerous questions remain unanswered, including, "why are some of the busiest airports not always in the most populous cities?", or ,"why are more half of the busiest international airports in Europe and North America, where only 20% of global population?"; or even: "Why does a 45-minute flight from Gaborone to Johannesburg still cost at least $200?". These are some of the questions we hope to uncover in our network analysis.

In this paper, we analyze the WAN using theoretical measures to investigate whether given geographical and/or geopolitical factors may substantially shape its organizational properties. We use

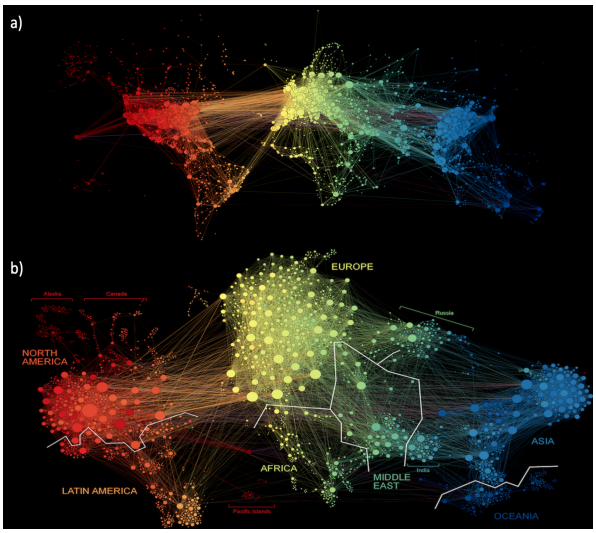

Fig. 1. Both figures show connections from over 3200 airports worldwide. (5). The color of the nodes corresponds to the longitude of the airport while the size signifies the number of routes associated with the airport. Figure b) is the result of the application of a force-directed layout algorithm on a graph of 3275 airports (37153 single routes – the weighted total is higher because many airlines take the same route), based on OpenFlights.org data. Naturally, network geography is not completely disrupted: the continents are mostly visible and regions are generally in their original position (with the exception of the Pacific islands that connect Asia and America – imagine this graph in three dimensions, with the Pacific Ocean behind). Major observations: India is more connected to the Middle East than to South and East Asia. The Russian cluster is very visible, connecting airports in Russia but also in many former Soviet republics. Latin America is clearly divided between a South cluster and a Central American cluster very connected with the U.S. Figure a) is the geographical layout of the same graph

data from the top 500 busiest airports, acquired from July 1, 2007 to July 30, 2008. To construct our WAN, we consider each airport as a node, and regard the existence of a flight connecting two airports as an unweighted network edge. We hypothesize that the WAN may exhibit community structure that is heavily influenced by geographical factors, whereby its communities may primarily co-localize in distinct world continents. We begin our analysis by using different centrality measures, in an attempt to rank different airports by flight traffic and overall importance and influence in context of the WAN. We also find that the WAN exhibits scale-free properties based nodal degree distribution which proves the existence of hubs. Scale-free networks are more robust against failure. By this we mean that the network is more likely to stay connected than a random network after the removal of randomly chosen nodes. However, they are also more vulnerable against non-random attacks. This means that the network quickly disintegrates when nodes are removed according to their degree. We then, apply the Louvain community detection algorithms to partition the WAN into modules, or communities. From further analysis of nodes in such communities, we quantify each city's global role, which we characterize as patterns of intercommunity and intracommunity connections. The features uncovered from this investigation could be used to design a new generation global airline network that may supersede the current in terms of efficiency and convenience owing to fewer transits and shorter average travel times from any starting point to any destination on earth. We expect our work to be beneficial to travelers and airline companies,enhancing communications and transportation worldwide, while also to contributing to the reduction of greenhouse emissions in the near future.

Table 1. A summary of all centrality metrics used to rank the busiest and most influential airports in the WAN. After considering the different measures used and weighting their outputs accordingly, Table 1 was produced to depict the top 5 most of important airports

Country	IATA Code	In-degree	Out-degree	Betweeness	Page Rank	Hub
Germany	FRA	217	217	14342.93	0.00838	0.0065
USA	JKF	217	218	13431.28	0.00755	0.00914
Netherlands	AMS	205	211	12892.75	0.00789	0.00613
France	CDG	195	200	10514.31	0.00756	0.00569
USA	ORD	192	194	9493.79	0.00655	0.00868

Results

Nodal Influence. In an effort to rank to airports based on their influence and significance to the WAN, different statistical network theory measurements were utilized . Firstly, the degree of each node was calculated. The nodal degree corresponds to the number of edges connected to each node. Since the WAN is a directed network, the in-degree and out-degree had to be calculated. The in-degree is analogous to the number of inbound flights from different airports while the out-degree corresponds to number of outbound flights. Frankfurt Airport in Germany and Kennedy International Airport in New York City were tied for the largest in-degree with 217 inbound flights while John F. Kennedy International Airport had the largest out-degree with 218 outbound flights. For comparison, the median out-degree was $35(IQR = 46)$ while the median in-degree was $36(IQR = 45)$.

The betweeness centrality of each node was also used to rank the importance of each airport. The betweeness centrality measures how often each node appears on a shortest path between any two nodes in the graph. In the WAN, this corresponds to how often a given passenger may need to fly through a specific airport in order to connect to another flight. Big, important airports usually have a large amount of inbound and outbound traffic. Once again, Frankfurt Airport had the largest betweeness centrality with 14343 while the median for all 500 aiports was $71.1(IQR = 266.2)$.

The next centrality measure used to classify the nodal importance was the pagerank centrality measure. Like eigenvector centrality, the pagerank can be considered as the "importance score" of a node. This importance score will always be a non-negative real number and all the scores will add to 1. Once again, Frankfurt Airport had the largest importance measure with a pagerank centrality of 0.0084 while the median pagerank was $0.0017(IQR = 0.0013)$.

Lastly, we classified nodal importance using the hub centrality measure. A hub is a node in network with a high-degree, thus having significantly larger number of edges or links in comparison to the rest of the nodes. The number of links (degrees) for a hub in a scale-free network is much higher than for the biggest node in a random network. A Matlab in-built function generated hub scores for each node similar to the pagerank function where the sum of all hubs scores is 1. The node with the highest hub score was John F Kennedy with 0.00914 while the median hub centrality measure was $0.0013(IQR = 0.0023)$.

Modularity. Modularity is the measure of strength of division of a network into individual modules (also called groups, clusters or communities) in a network. A network with dense connections within modules but sparse connections between other modules has high modularity. Modularity is often used in optimization methods for detecting community structure in networks. However, it has been shown that modularity suffers a resolution limit and, therefore, it is unable to detect small communities. Before detecting and identifying different communities in the network, we had to identify

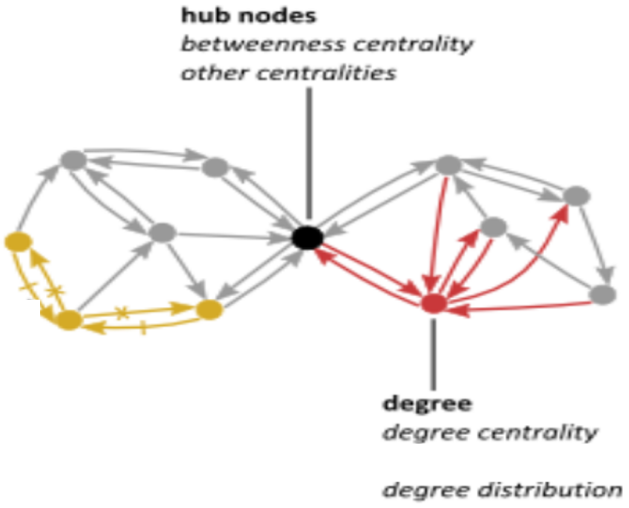

Fig. 2. An illustration of hub nodes and degrees in a network. Hubs usually have higher overall degrees than other nodes. In order to classify certain nodes as hubs, a threshold is set and any node that has a hub score that is greater than the threshold is classified as a hub.

the resolution parameter for the out community detection algorithm. To optimize for the right γ value, we ran the algorithm with varying γ values ranging from 0 to 3, evenly spaced by 0.1. After plotting the Q vs the γ, the point on the graph with the largest gradient was found to be $\gamma = 1$, which was the γ parameter to find the number of communities.

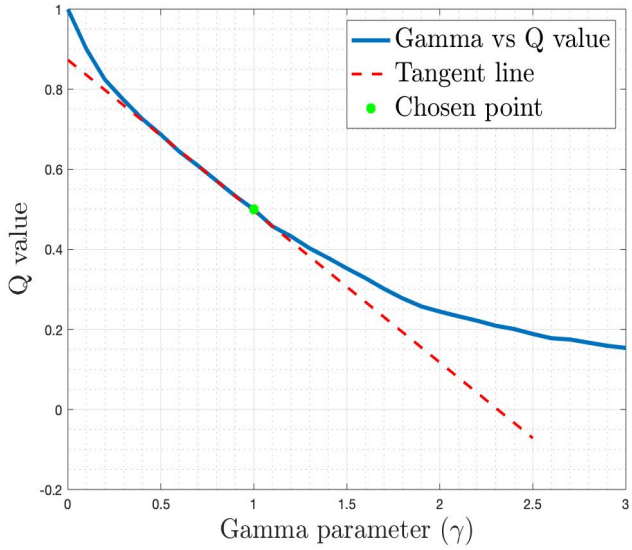

Fig. 3. Q vs γ graph. The point were $\frac{dy}{dx}$ is highest exists at the point were $\gamma = 1$. We can use this point in in the algorithm as the best balance between Q and γ

Once the γ resolution parameter was defined, we used the Louvain community detection algorithm to find the communities within the WAN. The Q value, the output from the Louvain community detection algorithm, is defined as a value in the range $[\frac{-1}{2}, 1]$ that measures the density of links inside communities compared to links between communities. A total of 4 communities were detected within the network as seen on figure 5. The algorithm was

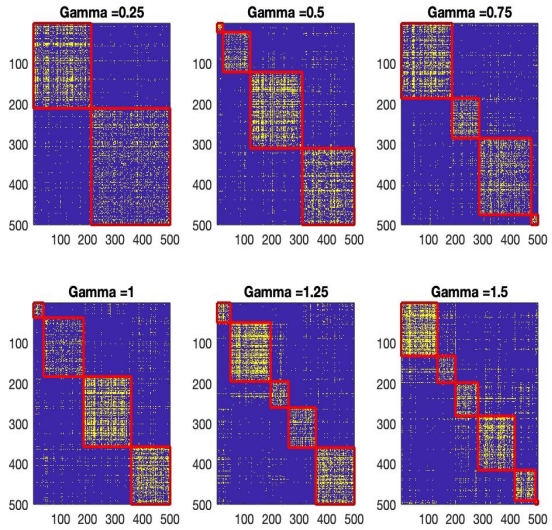

Fig. 4. Communities from different γ values ranging from 0.25 to 1.5 are shown.It was observed that as γ value increases, the number of communities also increases. However, as γ increases, the size of communities get smaller. This graph is mean to represent the importance in finding the correct γ parameter

computed a total of 10,000 times. The average Q value obtained from the algorithm was 0.4993 ± 0.0011.

Significance of Modularity. In order test if the average Q value obtained from the Louvain algorithm was statistically different from a random network, its significance was tested. A random directed network was generated with 500 nodes to compare to the real data-set. Random graphs were computed 10,000 times and their Q values were plotted on a histogram. The data proved to be normally distributed as seen on figure 6. The mean of the 10 000 randomly samples was $0.0245(\sigma = 6.256e - 6)$. A Z-score of 772.2 was calculated from the 10 000 samples resulting in a P-value that was $< .0001$. The formula used to attain the Z-score was:

$$z = \frac{x - \mu}{\sigma}$$

where x is the observed Q value from the real data while mu and σ are the mean and of the standard deviation of the 10 000 random samples.In other words, the average Q value of obtained from the real is much more extreme than all the other Q values obtained from the random networks. Since this P-value was less than a 0.05, it showed that the modularity of the real network was not due to random chance.

Scale Free network . The existence of hubs is the biggest difference between random networks and scale-free networks (11). Hubs in a network can be found by setting a threshold to the hub score, classifying all nodes above the threshold as hubs. We used an arbitrarily chosen threshold of 0.003 and found that 116 nodes that were hubs. We compared this to a random generated network with 500 nodes and found no nodes had scores above 0.0025 or scores below 0.0019. Remember that all the hubs scores in a network have to add up to 1. We generated the hubs scores from random networks, 10 000 times, and found that the average interquartile range was $1.2012e - 04$ while the real data had hub

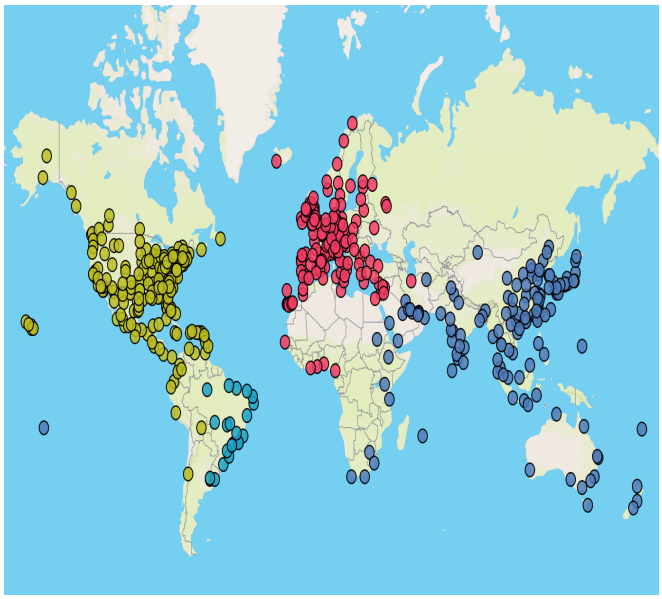

Fig. 5. Different clusters or communities of the WAN are represented by different colors. Each dot represents an airport location. The figure also shows how some of the communities overlap. One can notice how most of the communities are clustered in North America and Europe. It also appears as though South America and Africa have fewer airports and their continental communities overlap with the continents to their north.

score interquarlile range of 0.0023. To test the statistical significance, a similar test was the carried out on Q values to produce figure 6. It was found that the Z score was 333.9 resulting in a P-value that was $< .0001$. We concluded that the existence of hubs in the WAN was not due to random coincidence. In scale-free networks, a few nodes (hubs) have a high degree while the other nodes have a small number of links (11).

Both the in-distribution and the out-distribution were tested for skewness using the MATLAB function skewness. The out-degree had a positive skewness of 1.62 while the in degree also a had a positive skewness of 1.56. The positive value of skewness meant that the distribution had a positive skew thus resulting in the distribution being right skewed. The nodal degree distribution of the WAN was compared to a random directed network with 500 nodes as well. The random network had a slightly positive out-degree skewness of 0.0592 while the in-degree skewness was also barely greater than zero at 0.1197. This shows indicates that the skewness nodal degree distribution is not due random chance thus the network is organized displays a scale-free network property.

To test the robustness of the network, we tested the global efficiency and Mean first passage time (MFPT) then removed the 132 hubs before testing the WAN network again. Global efficiency is a measure of the efficiency of distant information transfer in a network and is defined as the inverse of the average characteristic path length between all nodes in the network(12). The Mean first passage time of network is the MFPT first-passage time of First-passage times over all source nodes in the network, which is a useful tool to analyze the behavior of random walks(13). The global efficiency of the WAN was 0.49 while the MFPT was $1.247 * 10^3$. After the 132 identified hubs were removed, global efficiency decreased to 0.34 while the MFPT increased to $6.6112 * 10^{14}$, nine orders of magnitude greater than the initial MFPT.

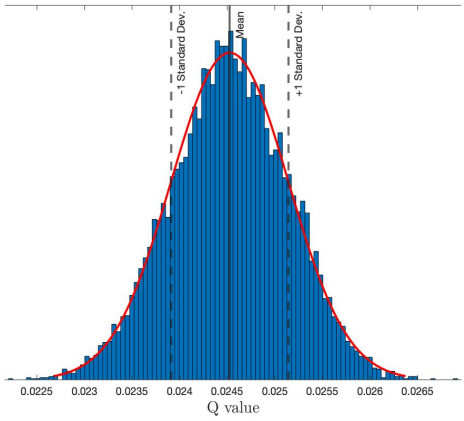

Fig. 6. The distribution of 10 000 iterations of randomly generated directed networks with 500 nodes. About 95% of the Q values were between 0.0233 and 0.0258. This proves that the average Q value obtained of 0.4993 was statistically different from any random data.

Discussion

The aim of this investigation was to analyze the structure of the WAN to uncover its topological features. We started on a nodal level and discovered how different nodes have different levels of importance and influence on the WAN. We used nodal degree, betweeness centrality, pagerank centrality and hub centrality to rank the importance of each node is in the network. This was analogous to uncovering the busiest and largest airports based on traffic routes and number passengers that pass through the airport. We found that many of the important airports are in Europe and North America, such as, John F Kennedy International Airport in USA and, Frankfurt International Airport in Germany. These appeared frequently among the top 2 in different centrality measures. Subsequently, we investigated the community structure of the WAN uncovering 4 distinct clusters of airports that were tightly interconnected within the clusters but weakly connected between each cluster. This confirmed our hypothesis of the existence of community structure. However, we discovered that these communities are not necessarily distributed by continent as seen on figure 5. Lastly, we assessed the network structure as a whole. We found that the WAN exhibits some scale-free network properties. This was not a surprise as many real world networks such as train routes, power grids and metabolic interactions also exhibit scale free network properties (11).

Airport Importance. A variety of centrality measures were used to rank airport importance since there is no definitive method of calculating the importance of a node in network. These metrics are descriptive of different characteristics of a node's centrality in network rather than indicating overall significance to the network. For example, the degree of a node, defined as the total number of relationships or links or edges involving that node, permits comparisons between network nodes. Nodes with higher degree values are more active than those with lower values. In contrast, a node with higher betweenness centrality would have more control over the network, because more information will pass through that node. Betweenness centrality is related to a network's connectivity, so much so that high betweenness vertices have the potential to disconnect graphs if removed (14).Both measures describe different forms of node activity and cannot independently unveil which

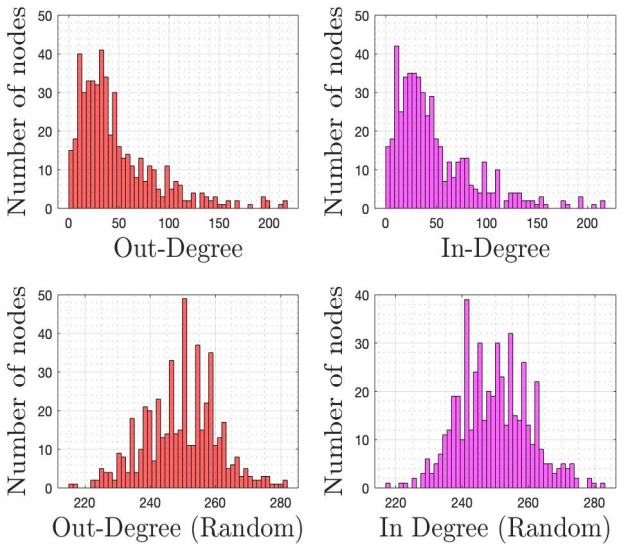

Fig. 7. This figure shows the distribution of the out-degree and the in-degree of the World Wide Network (WAN) compared to the nodal degree distribution of a random network. Both distributions of the WAN are skewed while the random directed network presents a normal distribution. A heavy tailed nodal distribution is on of the properties of a heavy scale free network as it shows there are a few nodes with high degrees.

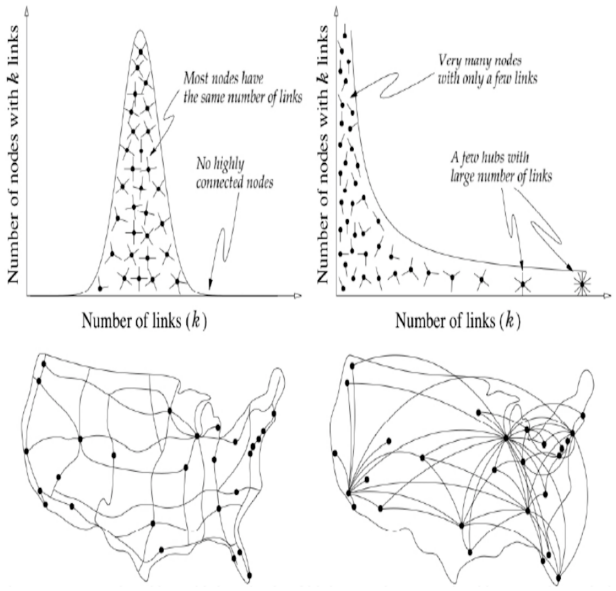

Fig. 8. The figure shows what air transportation network would be like in the USA using a network with normal degree distribution vs a network with scale free network.

nodes are the most important based on traffic or total number of nodes that would be affected if was removed. The page rank was developed by the Google founders when they were thinking about how to measure the importance of webpages using the hyperlink network structure of the web.The basic idea is that, pagerank will assign a score of importance to every single node. It assumes that important nodes are those that have many in-links from important pages or other important nodes. PageRank centrality can be also used on any type of network, for example, the web or social networks, but it really works better for networks that have directed edges which is the case for the WAN(15). We added hub scores after noticing the skewness in distribution of node degree since a hub is a node with a number of links that greatly exceeds the average. The rankings from these metrics were somewhat consistent with minor differences in rankings as seen on table 1. We averaged the scores across airports and found that the most important ones were Frankfurt International Airport and John F Kennedy International Airport. Given this information, we can allocate more resources to these airports as they benefit the entire WAN to increase efficiency. Such resources include security measures to protect them from attacks or upgraded health measures to decrease disease spreading. The hubs are also responsible for effective spreading of material or information in network. In an analysis of disease spreading or information flow, hubs are referred to as super-spreaders (16) . Super-spreaders may have a positive impact, such as effective information flow, but may also be devastating in a case of epidemic spreading such as H1N1 or Covid-19. The mathematical models such as the model of H1H1 Epidemic prediction may allow us to predict the spread of diseases based on human mobility networks, infectiousness, or social interactions among humans (16). Therefore, airports hubs are also important in the control of disease spreading.

Airport Communities. The Louvain community algorithm uncovered the 4 different airports clusters that are heavily connected within the clusters but but weakly connected between the different clusters. There are different kinds of community detection algorithms, but we chose to use the Louvain community algorithm due to its rapid convergence properties, high modularity and hierarchical partitioning.(17).Before applying the Louvain community detection algorithm, we had to optimize for the resolution parameter (γ) in the algorithm . This was done in order to ensure the correct trade-off between the number of modules detected and the size of each community. The relation between the γ parameter and size of communities detected can be observed in figure 4. This lead to 4 modules detected using a γ value of 1.

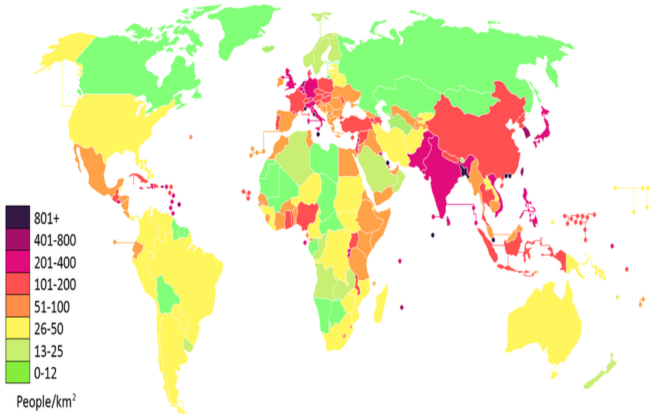

Fig. 9. This figure was generated with data from (18) and shows population density per country. Asia is most densely populated continent followed closely by Europe. Even though Africa has the second largest population by continent, it is still relatively sparsely populated. North and South America have similar population densities.

The distribution of the airport communities was largely by continent with North America and Europe having the largest clusters as seen on figure 2. This was a surprising discovery since Asia and Africa have the largest continental population accounting for approximately 77% of the global population combined. Regardless

of the regional population, this disparity in community structure may be due to economic activity. Maps of the global distribution of economic activity by Ghosh et al. correlate to the of density of clusters found on different continents (19). There some important questions that arise upon this discovery. Shouldn't more people equate to more transportation availability? Does the lack of WAN infrastructure limit economic activity growth or does the lack economic activity limit the growth of infrastructure.Assuming there are more people in a given region, should there not be more opportunities for safe and efficient travel. Perhaps the answer lies in population density rather than continental population. Examining both figure 5 and figure 9, it is evident that there is a correlation between density and the structure of the WAN, but it is still not the primary driving factor. If modularity of the WAN is not based on geography and populations dynamics, then perhaps it is based on geopolitics. Travel restrictions, tariffs and international organizations drive economic activity which affects number of airports and travel routes available in certain geographical regions. These undeserved geographical locations of the WAN have resulted in more expensive flights per unit of distance compared to the North America and Europe which further limit economic growth in those regions.

Impact of scale free network structure. The nodal degree distribution and the existence of hubs led us to believe that the WAN has scale free network structure (20). Scale free networks have "small world" properties in that they exhibit short typical path lengths which results in shorter average flight times. (21). However, the universality of scale-free networks remains controversial while the unclear standards of what counts as evidence for or against the scale-free hypothesis remains (20). Typically, scale free networks usually have the following properties: (i) all node are not equal and some are highly connected compared to others, (ii) robust against accidental failures but note coordinated attacks, (iii) linkages follow the power law (20). Table 7 shows the WAN has a few nodes with a large number of edges while the most have less than 50 links. This is usually because older node are generally more fit and will get more links over time (20). Preferential attachment plays a role as new nodes will prefer to connect to older nodes thus employing "the rich get richer" concept. Preferential attachment is linear as older nodes have a fixed higher probability. Another case would be the "Winner takes all approach" which would result in star topology of the network.

Robustness to localized failures is a general strength of many network constructions (22). Because of the scale free nature of the WAN, issues such as airport weather closures are more easily resolved because they are more flexible and can utilize alternate links in their network. Some of this flexibility is inherent in a multi-hub operation, where passengers can be re-routed away from problem areas. However, without an ability to also adjust resources across routes, the flexibility of extensive networks operating at near-full capacity is limited (21). The downside to such constructions is vulnerability to disruption. Networks with hubs that have a high probability of experiencing problems also have a high probability of proliferating those problems across the entire network (21). Particularly in the area of air traffic management, where the hubs are largely constructions of the operational control mechanisms (e.g. multiple aircraft to a controller), diversification of the control task could lower the vulnerability to a disruption (21).

Limitations and future directions. The biggest shortcoming of this study is the small number of nodes relative to the true number of airports in the WAN. Even though the data used accounted for 95% of the global airline traffic, only 500 nodes were included out of the total 42 000 airports globally.Therefore, the conclusions made from this investigation assume the WAN is shaped by only 500 of the 40 000+ available airports. Analyzing all 40 000 nodes is computationally taxing but important if we aim to achieve more accurate conclusions. Future studies can include more extensive datasets to gain a more holistic picture of the topologically structure of the WAN.

Furthermore, the data used was acquired from 2008 which is more than a decade ago. The WAN is dynamic in nature as new airports are built to support growing populations or international events. For example, Hamad International Airport in Doha, Qatar has grown over the past decade to became the worlds best rated airport. This may be in preparation to host the FIFA world cup in 2022 which will see a large influx of passengers into the country during that period. This study provides a framework for future studies to investigate the driving factors that affect the growth of the WAN. Our study has uncovered a discrepancy in airport placement in primarily in Africa. Future studies should investigate further whether population dynamics or economic activity led to this discrepancy. These works may also use data acquired from different decades to correlate their finding to population dynamics as well as economic activity.

Conclusion

The World Airline Network (WAN) is an infrastructure that aims to reduce the geographical gap between global societies. Using publicly available data, we have presented a full topological analysis of the WAN. We ranked the most important airports using different centrality metrics and uncovered the existence of hubs. Most hubs were situated in North America and Europe. We characterized the community structure of WAN using the Louvain algorithm and discovered 4 different modules with North America and Europe once again being the largest communities. The existence of hubs, along with other properties,led to the classification of the WAN as a scale free network structure. There are several shortcomings and advantages of scale free network organisation which were discussed. Lastly, we briefly discussed possible future directions provided by the points discussed in this study

Methods and Materials

Airline Data. The data was collected from scheduled flight data that had been provided over the course of about a year (July 1, 2007 to July 30, 2008) [11]. According to the data, 1,341,615 flights were recorded by OAG Avian solutions is a provider of digital flight information, intelligence and analytics for airports, airlines and travel tech companies.[16]. It consisted of data from the top 500 busiest airports across the world. Although this was not an extensive list of all the possible flights in those airports at the time, it included approximately 99 % of all commercial flights across the 500 major airports. A complete list of the airports can be found in Table 2. The data included the cities which the flights originated and were destined to. In this network, a node is represented as an airport and an edge is described as a flight from one airport to another. The direction of that the flight indicates the direction of the edges thus indicating this is a directed network. The original data-set can be found at http://www.biological-networks.org/.

Network centrality measures. Matlab functions were used to find the degrees of each node.The 'out-degree', and 'in-degree'

centrality types are based on the number of edges connecting to each node. Using the inbuilt the Matlab function "centrality" with either 'outdegree', or 'indegree' as arguments of the function, the degrees of each nodes was attained from the adjacency matrix . It is important to note that 'indegree' corresponds to the number of incoming edges to each node and a self-loop counts as one incoming edge while the 'outdegree' corresponds to number of outgoing edges from each node. These functions find the degrees of each node and which then sorted by their nodal degree to find the find the busiest airports.The out-degree is the number of outgoing edges emanating from a node:

$$k_{out}^{i} = \sum_{j} a_{ji}, \qquad [1]$$

and the in-degree is the number of incoming edges onto a node is:

$$k_{in}^{i} = \sum_{j} a_{ji}, \qquad [2]$$

The betweeness centrality was also calculated using inbuilt Matlab functions. The betweenness centrality measures how often each node appears on the shortest path between two nodes in the graph. Since there can be several shortest paths between two graph nodes s and t, the centrality of node u is:

$$c(u) = \sum_{s,t \neq u} \frac{n_{st}}{N_{st}}, \qquad [3]$$

$N_{st}(u)$ is the number of shortest paths from s to that pass through node u, and N_{st} is the total number of shortest paths from s to t.

The 'pagerank' centrality type results from a random walk of the network and was measured using an inbuilt Matlab function. At each node in the graph, the next node is chosen with probability 'FollowProbability' from the set of successors of the current node. Otherwise, or when a node has no successors, the next node is chosen from all nodes. The pagerank centrality score is the average time spent at each node during the random walk.There are three distinct factors that determine the pagerank of a node: (i) the number of links it receives, (ii) the link propensity of the linkers, and (iii) the centrality of the linkers. The first factor is not surprising as the more edges/links a node attracts, the more important it is perceived. Reasonably, the value of the endorsement depreciates proportionally to the number of edges given out by the endorsing node. Finally, not all nodes are created equal: links from important vertices are more valuable than those from obscure ones.

Let $A = (a_{i,j})$ be the adjacency matrix of a directed graph. The PageRank centrality x_i of node i is given by:

$$x_i = \alpha \sum_{k} \frac{a_{k,i}}{d_k} x_k + \beta$$

where α and β are constants and d_k is the out-degree of node k if such degree is positive, or $d_k = 1$ if the out-degree of k is null. In matrix form we have:

$$x = \alpha x D^{-1} A + \beta$$

where β is now a vector whose elements are all equal a given positive constant and D^{-1} is a diagonal matrix with i-th diagonal element equal to $1/d_i$. Notice that, as seen for Katz centrality, pagerank is determined by an endogenous component that takes into consideration the network topology and by an exogenous component that is independent of the network structure. It follows that x can be computed as:

$$x = \beta(I - \alpha D^{-1} A)^{-1}$$

The 'hubs' and centrality scores are two linked centrality measures that are recursive. The hubs score of a node is the sum of the authorities scores of all its successors. An in-built hub-centrality function based on the Hub Update Rule and Authority Update Rule was used to calculate all the hub scores. To begin the ranking, it lets $auth(p) = 1$ and $hub(p) = 1$ for each page p. In order to calculate the hub/authority scores of each node, repeated iterations of the Authority Update Rule and the Hub Update Rule are applied. A k-step application of the Hub-Authority algorithm entails applying for k times first the Authority Update Rule and then the Hub Update Rule. The Authority Update Rule states that for each p, we update $auth(p)$ to

$$auth(p) = \sum_{q \in P_{to}} hub(q)$$

where P_{to} is all pages which link to page p. That is, a page's authority score is the sum of all the hub scores of pages that point to it.The Hub Update Rule states that for each p, we update $hub(p)$ to

$$hub(p) = \sum_{q \in P_{from}} auth(q)$$

where P_{from} is all pages which link to page p. That is, a page's hub score is the sum of all the authority scores of pages it points to.

Skewness. Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution. A normal distribution has a skew of zero, while a lognormal distribution, for example, would exhibit some degree of right-skew. An in-built Maltab function was used to calculate skewness which followed the basic formula defined as:

$$s = \frac{E(x-\mu)^3}{\sigma^3} \qquad [4]$$

where μ is the mean of x, σ is the standard deviation of x, and $E(t)$ represents the expected value of the quantity t.

Random Network Generation. The random network used for comparison was generated by randomly shuffling the 500x500 adjacency matrix that provided the flight data. Its output was another 500x500 a directed adjacency matrix that had rearranged the number of edges connected to each node. Due to this random permutation of elements in the matrix, the edges which connected certain nodes originally also changed. The function was ran 10 000 times to generate 10 000 random samples to compare to the real data. There were no self-loops or no double edges in the random graphs.

Community detection. The Louvain community detection algorithm was executed in Matlab. This function is a fast and accurate multi-iterative generalization of community detection algorithm. It allows us to optimize other objective functions and includes built-in Potts-model Hamiltonian and also allows for custom objective-function matrices [16].The inputs of the Louvain algorithm include

the adjacency matrix containing WAN data and the variable γ. γ is a resolution parameter that scales how small or large the communities should be once detected. A value of $\gamma > 1$ detects smaller modules or communities while a $\leq \gamma < 1$ value detects larger modules. The default γ value is 1. 31 different values of γ were tested starting with 0 increased by 0.1 until 3. Figure 4 shows how changing the γ parameter affected the number of communities detected as well as the number of nodes within a community.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \qquad [5]$$

where:

A_{ij} represents the edge weight between nodes i and j;

k_i and k_j are the sum of the weights of the edges attached to nodes i and j, respectively;

m is the sum of all of the edge weights in the graph;

c_i and c_j are the communities of the nodes;

δ is Kronecker delta function $\delta(x, y) = 1$ if $x = y$, otherwise).

Global Effeciency. Global efficiency is a measure of the efficiency of distant information transfer in a network and is defined as the inverse of the average characteristic path length between all nodes in the network (23). A Matalab function was written in order to calculate the global efficiency of the network. We began by calculating the shortest number of steps required to go from node i to every other network node was computed. This was done separately for each and every node in the network, and the average number of shortest steps to all other network nodes was computed separately for each node. The inverse of the average number of shortest steps for each node was then summed across all network nodes and this summed quantity is normalized by taking into account the total possible number of connections that could exist in the network. Formally, global efficiency is calculated as:

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in V} \frac{1}{d_{ij}}, \qquad [6]$$

where $N = |V|$ denotes the number of nodes in the network.

Appendices.

References.

1 M Zanin, F Lillo, Modelling the air transport with complex networks: A short review. *The Eur. Phys. J. Special Top.* **215**, 5–21 (2013).
2 ICAO, The economic social benefits of air transport (https://www.icao.int/Meetings/wrdss2011/Documents/Forms/AllItems.aspx?RootFolder=/meetings/wrdss2011/documents/jointworkshop2005&FolderCTID=0x0120) (2011).
3 (year?).
4 E Mazareanu, Passenger air traffic each year (2021).
5 PM GRANDJEAN, Connected world: Untangling the air traffic network (2016).
6 W Guo, et al., Global air transport complex network: multi-scale analysis. *SN Appl. Sci.* **1**, 1–14 (2019).
7 T Magnanti, L., and wong r. *T.: Netw. Des. Transp. Planning: Model. Algorithm Transp. Sci.* **18** (1984).
8 FA Stillman, A Soong, C Kleb, A Grant, A Navas-Acien, A review of smoking policies in airports around the world. *Tob. Control.* **24**, 528–531 (2015).
9 R Guimera, M Sales-Pardo, LA Amaral, Classes of complex networks defined by role-to-role connectivity profiles. *Nat. physics* **3**, 63–69 (2007).
10 R Guimera, S Mossa, A Turtschi, LN Amaral, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci.* **102**, 7794–7799 (2005).
11 M PÓSFAI, et al., Network science. (year?).
12 V Latora, M Marchiori, Efficient behavior of small-world networks. *Phys. review letters* **87**, 198701 (2001).
13 Y Lin, Z Zhang, Mean first-passage time for maximal-entropy random walks in complex networks. *Sci. reports* **4**, 1–7 (2014).
14 ME Newman, A measure of betweenness centrality based on random walks. *Soc. networks* **27**, 39–54 (2005).
15 D Romero, Applied social network analysis in python (2021).
16 D Balcan, et al., Seasonal transmission potential and activity peaks of the new influenza a (h1n1): a monte carlo likelihood analysis based on human mobility. *BMC medicine* **7**, 1–12 (2009).
17 X Que, F Checconi, F Petrini, JA Gunnels, Scalable community detection with the louvain algorithm in *2015 IEEE International Parallel and Distributed Processing Symposium*. pp. 28–37 (2015).
18 cartoMission, Population 2018 – cartomission (https://cartomission.com/2018/01/17/population-2018/) (2018) (Accessed on 11/30/2021).
19 T Ghosh, et al., Shedding light on the global distribution of economic activity. *The Open Geogr. J.* **3** (2010).
20 AD Broido, A Clauset, Scale-free networks are rare. *Nat. communications* **10**, 1–10 (2019).
21 S Conway, Scale-free networks and commercial air carrier transportation in the united states in *24th congress of the international council of the aeronautical sciences*. Vol. 29, (2004).
22 J Zhao, K Xu, Enhancing the robustness of scale-free networks. *J. Phys. A: Math. Theor.* **42**, 195003 (2009).
23 ML Stanley, et al., Changes in brain network efficiency and working memory performance in aging. *PLoS One* **10**, e0123950 (2015).
24 Airline and location code search (year?).

Table 2. This table contains the complete list of IATA codes of each airport used in this study.IATA codes are three-letter geocode designating many airports and metropolitan areas around the world, defined by the International Air Transport Association (IATA) (24). Along with the IATA, the ISO country codes as well as the continent of each airport are also given.

Continent	Iso Country	IATA code	Continent	Iso Country	IATA code	CContinent	Iso Country	IATA code	Continent	Iso Country	IATA code	Continent	Iso Country	IATA code
AF	CI	ABJ	NA	US	COS	NA	US	HNL	AS	MV	MLE	NA	US	SFO
NA	US	ABQ	EU	DK	CPH	NA	US	HOU	NA	US	MLI	AS	VN	SGN
EU	GB	ABZ			CPQ	NA	US	HPN	EU	GB	MME	AS	CN	SHA
NA	MX	ACA	AF	ZA	CPT	AS	CN	HRB	AS	PH	MNL	AS	CN	SHE
AF	GH	ACC	NA	US	CRP	NA	US	HRL	EU	FR	MPL	AS	AE	SHJ
EU	ES	ACE	AS	CN	CSX	NA	US	HSV	EU	FR	MRS	AS	SG	SIN
AS	TR	ADB	EU	IT	CTA	AS	IN	HYD	AF	MU	MRU	NA	US	SJC
AF	ET	ADD	AS	JP	CTS	NA	US	IAD	NA	US	MSN	NA	MX	SJD
OC	AU	ADL	AS	CN	CTU	NA	US	IAH	NA	US	MSP	NA	CR	SJO
SA	AR	AEP	NA	MX	CUL	EU	ES	IBZ	NA	US	MSY	NA	PR	SJU
EU	ES	AGP	NA	MX	CUN	AS	KR	ICN	NA	MX	MTY	EU	GR	SKG
OC	NZ	AKL	NA	MX	CUU	NA	US	ICT	EU	DE	MUC	NA	US	SLC
NA	US	ALB	NA	US	CVG	NA	US	IND	SA	UY	MVD	SA	BR	SLZ
EU	ES	ALC	SA	BR	CWB	AS	PK	ISB	EU	IT	MXP	NA	US	SMF
AF	DZ	ALG	EU	GB	CWL	AS	JP	ISG	AS	JP	MYJ	NA	US	SNA
NA	US	AMA	AS	BD	DAC	NA	US	ISP	NA	US	MYR	EU	IE	SNN
AS	IN	AMD	NA	US	DAL	AS	TR	IST	AS	TW	MZG	EU	BG	SOF
AS	JO	AMM	AS	SY	DAM	AS	JP	ITM	NA	MX	MZT	EU	GB	SOU
EU	NL	AMS	AF	TZ	DAR	NA	US	ITO	OC	FJ	NAN	NA	US	SRQ
NA	US	ANC	NA	US	DAY	NA	US	JAN	EU	IT	NAP	SA	BR	SSA
NA	AG	ANU	NA	US	DCA	NA	US	JAX	NA	BS	NAS	NA	US	STL
EU	SE	ARN	AS	IN	DEL	AS	SA	JED	SA	BR	NAT	EU	GB	STN
EU	GR	ATH	NA	US	DEN	EU	JE	JER	AF	KE	NBO	EU	DE	STR
NA	US	ATL	NA	US	DFW	NA	US	JFK	EU	FR	NCE	NA	VI	STT
NA	AW	AUA	AF	SN	DKR	AF	ZA	JNB	EU	GB	NCL	AS	ID	SUB
AS	AE	AUH	AF	CM	DLA	NA	US	JNU	AS	CN	NGB	EU	IT	SUF
NA	US	AUS	AS	CN	DLC	EU	UA	KBP	AS	JP	NGO	EU	NO	SVG
AS	TR	AYT	EU	RU	DME	AS	MY	KCH	AS	JP	NGS	EU	RU	SVO
AS	BH	BAH	AS	TH	DMK	EU	IS	KEF	AS	CN	NKG	EU	ES	SVQ
EU	ES	BCN	AS	SA	DMM	AS	TW	KHH	AS	CN	NNG	EU	FR	SXB
NA	US	BDL	AS	QA	DOH	AS	PK	KHI	AS	JP	NRT			
EU	RS	BEG	AS	ID	DPS	NA	JM	KIN	EU	FR	NTE			
SA	BR	BEL	EU	DE	DRS	AS	JP	KIX	EU	DE	NUE			
AS	LB	BEY	OC	AU	DRW	AS	CN	KMG	NA	US	OAK			
EU	GB	BFS	NA	US	DSM	AS	JP	KMI	NA	US	OGG			
NA	BB	BGI	NA	US	DTW	AS	JP	KMJ	AS	JP	OKA			
EU	NO	BGO	EU	IE	DUB	AS	JP	KMQ	NA	US	OKC			
EU	IT	BGY	AF	ZA	DUR	NA	US	KOA	NA	US	OMA			
EU	GB	BHD	EU	DE	DUS	AS	JP	KOJ	NA	US	ONT			
NA	US	BHM	AS	AE	DXB	EU	PL	KRK	OC	AU	OOL			
EU	GB	BHX	EU	GB	EDI	AF	SD	KRT	EU	PT	OPO			
NA	US	BIL	NA	US	ELP	NA	US	KTN	NA	US	ORD			
EU	ES	BIO	EU	GB	EMA	AS	MY	KUL	NA	US	ORF			
NA	MX	BJX	AS	TR	ESB	AS	CN	KWE	EU	IE	ORK			
AS	MY	BKI	NA	US	EUG	AS	KW	KWI	EU	FR	ORY			
AS	TH	BKK	NA	US	EWR	AS	CN	KWL	EU	NO	OSL			
EU	DK	BLL	SA	AR	EZE	NA	US	LAS	EU	RO	OTP			
EU	IT	BLQ	NA	US	FAI	NA	US	LAX	EU	ES	OVD			
AS	IN	BLR	EU	PT	FAO	EU	GB	LBA	NA	US	PBI			
NA	US	BNA	NA	US	FAR	NA	US	LBB	NA	US	PDX			
OC	AU	BNE	NA	US	FAT	AS	CY	LCA	AS	CN	PEK			
EU	FR	BOD	EU	IT	FCO	EU	ES	LCG	AS	MY	PEN			
SA	CO	BOG	NA	MQ	FDF	EU	GB	LCY	OC	AU	PER			
NA	US	BOI	NA	US	FLL	EU	RU	LED	NA	US	PHL			
AS	IN	BOM	SA	BR	FLN	EU	DE	LEJ	NA	US	PHX			
EU	NO	BOO	EU	IT	FLR	NA	US	LEX	NA	US	PIT			
NA	US	BOS	EU	PT	FNC	NA	US	LGA	AF	ZA	PLZ			
EU	DE	BRE	NA	US	FNT	NA	US	LGB	EU	ES	PMI			
EU	IT	BRI	AS	CN	FOC	EU	GB	LGW	EU	IT	PMO			
EU	GB	BRS	SA	BR	FOR	AS	PK	LHE	AS	KH	PNH			
EU	BE	BRU	EU	DE	FRA	EU	GB	LHR	AS	IN	PNQ			
SA	BR	BSB	NA	US	FSD	AS	CN	LHW	NA	US	PNS			
EU	FR	BSL	EU	ES	FUE	NA	US	LIH	SA	BR	POA			
NA	US	BTR	AS	JP	FUK	SA	PE	LIM	NA	TT	POS			
NA	US	BTV	AS	IN	GAU	EU	IT	LIN	OC	PF	PPT			
EU	HU	BUD	NA	MX	GDL	EU	PT	LIS	EU	CZ	PRG			
NA	US	BUF	NA	US	GEG	NA	US	LIT	EU	IT	PSA			
NA	US	BUR	SA	BR	GIG	EU	SI	LJU	NA	US	PSP			
NA	US	BWI	EU	GB	GLA	AF	NG	LOS	NA	GP	PTP			
NA	US	BZN	AS	KR	GMP	EU	ES	LPA	NA	PA	PTY			
NA	US	CAE	EU	IT	GOA	EU	GB	LPL	NA	DO	PUJ			
AF	EG	CAI	AS	IN	GOI	EU	GB	LTN	AS	KR	PUS			
AS	CN	CAN	EU	SE	GOT	EU	LU	LUX	NA	US	PVD			
OC	AU	CBR	NA	US	GRB	EU	FR	LYS	AS	CN	PVG			
SA	VE	CCS	EU	ES	GRO	AS	IN	MAA	NA	MX	PVR			
AS	IN	CCU	NA	US	GRR	EU	ES	MAD	NA	US	PWM			
EU	FR	CDG	SA	BR	GRU	NA	US	MAF	AF	MA	RAK			
AS	PH	CEB	EU	AT	GRZ	EU	ES	MAH	NA	US	RDU			
SA	BR	CGB	NA	US	GSO	EU	GB	MAN	SA	BR	REC			
SA	BR	CGH	NA	US	GSP	SA	BR	MAO	AS	MM	RGN			
AS	ID	CGK	NA	GT	GUA	NA	JM	MBJ	NA	US	RIC			
EU	DE	CGN	OC	GU	GUM	NA	US	MCI	EU	LV	RIX			
AS	CN	CGO	EU	CH	GVA	NA	US	MCO	NA	US	RNO			
AS	CN	CGQ	SA	EC	GYE	AS	OM	MCT	NA	US	ROC			
OC	NZ	CHC	SA	BR	GYN	SA	CO	MDE	NA	US	RSW			
NA	US	CHS	EU	DE	HAJ	NA	US	MDT	AS	SA	RUH			
NA	US	CID	AS	CN	HAK	NA	US	MDW	AS	YE	SAH			
NA	MX	CJS	EU	DE	HAM	OC	AU	MEL	NA	SV	SAL			
AS	KR	CJU	AS	VN	HAN	NA	US	MEM	NA	US	SAN			
AS	CN	CKG	NA	CU	HAV	AS	ID	MES	NA	US	SAT			
NA	US	CLE	OC	AU	HBA	NA	MX	MEX	NA	US	SAV			
SA	CO	CLO	EU	FI	HEL	NA	US	MFE	NA	US	SBA			
NA	US	CLT	EU	GR	HER	AS	MO	MFM	NA	US	SBN			
AS	LK	CMB	AS	CN	HGH	NA	NI	MGA	SA	CL	SCL			
NA	US	CMH	AS	JP	HIJ	AS	IR	MHD	EU	ES	SCQ			
AF	MA	CMN	AS	JP	HKD	NA	US	MHT	NA	US	SDF			
SA	BR	CNF	AS	HK	HKG	NA	US	MIA	AS	JP	SDJ			
OC	AU	CNS	AS	TH	HKT	NA	MX	MID	NA	DO	SDQ			
AS	TH	CNX	NA	MX	HMO	NA	US	MKE	SA	BR	SDU			
AS	IN	COK	AS	JP	HND	EU	MT	MLA	NA	US	SEA			