Impact of COVID-19 Pandemic on Global Commercial Flight Network

Besher Mahrouka
York University
besherm@my.yorku.ca

Juan Leaniz
York University
jleaniz@my.yorku.ca

1 ABSTRACT

The goal of the project was to study the effect of COVID-19 on flight network. In this project, two flight networks were generated. One pre-COVID-19 and the other during COVID-19. Those flight networks were used to check ,which airport, and airline saw the highest reduction in air travel. The overall results showed that there was a decreased in number of flights during COVID-19. This result was obtained by using different flight network analysis techniques such as page ranking, HITS algorithm, degree distribution,clustering coefficient,,short path ,and link prediction by using betweenness algorithm.

2 MOTIVATION

Commercial air travel is a multi-billion dollar industry that provides employment for millions of people throughout the globe. In 2020, the COVID-19 pandemic has had an unprecedented impact on society and the way we conduct business. The global travel restrictions and lockdown rules that have been implemented in every country to prevent the spread of the virus have had a negative impact on almost every business, but no industry has been as severely impacted as the airline industry.

According to recent data published by the International Civil Aviation Organization (ICAO) [1], the airline industry is on track to sustain approximately between $386 to $399 billion United State Dollars (USD) in potential losses due to a reduction in air travel. These loses will undoubtedly have an effect on the routes each airline will continue to service, and which routes will be discontinued altogether due to the significant reduction in air travel.

The main objective of this project is to utilize available worldwide commercial flight data from 2019 and 2020 to investigate how the global flight network has been impacted by the COVID-19 pandemic. Performing network analysis on the global flight network will allow to better understand specific properties of the network, pre and post COVID-19. The outcome of the analysis in this project enables us to answer some important questions of interest. In particular, which airport saw the highest reduction in air travel?, which airlines saw the highest reduction in air travel? which paths (or links) have disappeared from the flight network due to the COVID-19 pandemic?, would it be possible to predict any future changes in the paths or links across the flight network, and if so, which links would be affected?. Have certain airports become more or less influential since the enactment of COVID-19 travel restrictions?.

3 PROBLEM DEFINITION

As it was mentioned in the previous section, the project is about studying the effects of COVID-19 on the flight network. For example, which airport, and airline saw the highest reduction in air travel. This can be achieved by calculating particular properties of the flight network, such as node degree, degree distribution, path lengths, average path length, average clustering coefficient and Node Centrality. The constrain in this project was using data that was balanced just to increase the accuracy of network analysis. For example, 7 months of data was used for pre-COVID-19, and 7 months of data was used during COVID-19. The restriction of the project was the data was not big enough ,because COVID-19 is still a recent problem. This might affect the accuracy of the analysis.

4 RELATED WORK

In a recent work,[2], the authors describe how to apply Markov Clustering algorithm based on the idea of random walks to analyse the global flight network with Pagerank scoring to rank the airports in each cluster. One of our research questions relates to finding how the importance of airports was affected during the pandemic. Pagerank provides a way to compute a score for each node's importance. We computed the Pagerank score for each airport in two separate graphs that represent the flight network prior to COVID-19 and during COVID-19 and analyzed their differences. A journal article published earlier in June 2020 by Suzunura et al. [3] analyzed an earlier version of the Opensky network data to identify flight patterns and densities around the world. The researchers observed relationships between the number of flights and an increase in COVID-19 infections, The research concludes that there was a 51% decrease in the global flight network density during the COVID-19 period in 2020. In our research, we build upon their approach by analyzing a newer version of the OpenSky netowrk dataset. We analyze global flight data and continue researching interesting changes in the properties of the global flight network caused by the COVID-19 pandemic.

5 METHODOLOGY

In this paper we utilize a free, crowd-sourced dataset provided by the OpenSky Network [4]. The OpenSky Network has compiled air traffic data from thousands of sensors across the globe. These sensors obtain air traffic data from the Automatic Dependent Surveillance-Broadcast (ADS-B) and Mode S. These technologies are used by aircraft to provide information to ground stations over the publicly accessible 1090 MHz radio frequency channel. Our dataset is comprised of air traffic data from January 2019 to August 2020 ,which has been made available to the public for free as a series of GZip compressed CSV files.

5.1 Dataset description

The dataset provided by the OpenSky Network [4] contains information observed from thousands of flights between January 2019 and August 2020. The total size of the dataset is approximately

2.4 GB, with 36,430,262 rows. Each row in the dataset represents a flight and provides the following items:

- **callsign**: the identifier of the flight displayed on ATC screens (usually the first three letters are reserved for an airline: AFR for Air France, DLH for Lufthansa, etc.)
- **number**: the commercial number of the flight, when available (the matching with the callsign comes from public open API)
- **icao24**: the transponder unique identification number.
- **registration**: the aircraft tail number (when available)
- **typecode**: the aircraft model type (when available)
- **origin**: a four letter code for the origin airport of the flight (when available)
- **destination**: a four letter code for the destination airport of the flight (when available)
- **firstseen**: the UTC timestamp of the first message received by the OpenSky Network.
- **lastseen**: the UTC timestamp of the last message received by the OpenSky Network.
- **day**: the UTC day of the last message received by the OpenSky Network.
- **latitude_1, longitude_1, altitude_1**: the first detected position of the aircraft.
- **latitude_2, longitude_2, altitude_2**: the last detected position of the aircraft.

The dataset can be downloaded freely at https://zenodo.org/record/4088202#.X6gqWmhKguU

5.2 Representing the Air Transportation Network

More precisely, major airports in the world are viewed as hubs that connect directly with hundreds of airports without intermediate routes. Other airports may by inter-connected via one or more intermediary airport. The global air transportation system can be represented as a network, in which nodes denote airports and an edge will be created if a direct flight exists between two airports. This project aims to analyze the network using data prior to the COVID-19 pandemic and compare it to a network built using data after COVID-19 travel restrictions were enacted to identify potential changes to the network's structure. Structural changes may relate to changes in the rank or importance of particular airports (i.e. nodes), changes in paths between airports (i.e. new or missing routes/links between airports), or airline-specific changes like the complete disappearance of an airline impacted by COVID-19 travel restrictions.

The network was built using the NetworkX Python library. Prior to creating the network with NetworkX, the data was processed with the Pandas library to generate the required edges and nodes. The node ,edge ,edge attribute can be represented as follows:

```
origin destination {'attribute': x}
```

where the 'attribute' field represents an object containing additional data related to the edge. The origin and destination fields represent the origin and destination airports (i.e. nodes).

The edge list can be subsequently imported into a Graph with NetworkX as follows:

```python
import networkx as nx
# Given a Pandas dataframe 'df' containing the edge list
G = nx.Graph()
G_flights = nx.from_pandas_edgelist(df,source='origin',
    target='destination',edge_attr='callsign',
    create_using=G)
# do something with graph 'G_flights'
```

Listing 1: Generating flight netowrk with NetowrkX

6 EVALUATION

In this paper, we build upon the research data provided by OpenSky to investigate the global air transpiration network for the period between June 2019 and September 2020. Our initial approach to perform the network analysis was to create an unweighted graph, in which nodes are the origin and destination airports, and an edge between the nodes will exist if and only if there exists a direct flight between the airports. We analyzed structural properties-node degrees, shortest path lengths, graph diameter, average shortest path length and clustering coefficient-to better understand the network.

6.1 Number of flights

The flight information was processed from the OpenSky dataset's CSV files. We extracted a sub-string of the callsign (i.e. flight identifier described in Section 4.1) code, which was then used as an edge attribute when creating the graph. The graph edge connects two nodes, the origin and destination, which represent the departure and arrival airports a flight used for its trip. Finally, after extracting the required information from the dataset, the number of flights for each company was calculated and sorted in descending order. The results for the top 30 airlines were plotted on an x-y axis graph for visualization purposes.

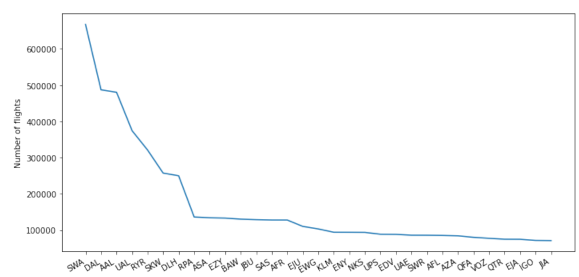


Figure 1: Number of flights before COVID-19.

We observed from Figure 1 that Southwest Airlines (SWA) has the highest number of flights, followed by Delta Airlines (DAL), American Airlines (AAL), United Airlines (UAL), Ryanair Ltd (RYR), SkyWest Airlines (SKW), Deutsche Lufthansa AG (DLH) and the last three companies are: Netjets Aviation (EJA), IndiGo (IGO) and Jetstream International Airlines (JIA).

By comparing the number of flights for each airline before the pandemic to the number of flights during the pandemic, we observed that the number of flights was significantly reduced for each airline. Figure 4 illustrates this observation.

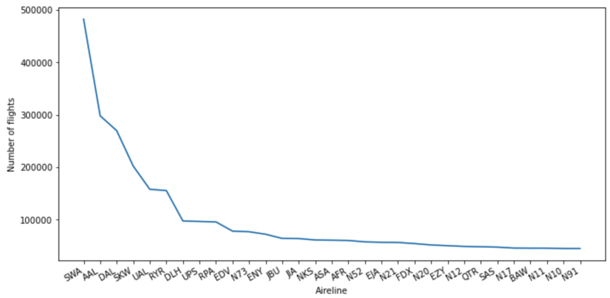


Figure 2: Number of flights during COVID-19.

Figure 3 shows that same companies are still on the top of the list: (SAW), then (AAL), (DAL) and (SKW) but this time with lower number of flights (more than 600.000 flights before COVID-19 and less than 500.000 flights since the flight restrictions were implemented).

6.2 Shortest Path Lengths

To determine the shortest path lengths we used Dijkstra's algorithm which is provided by the NetworkX Python library. Because our graph is unweighted, the *shortest_path_length* function assigns each edge a weight/cost of 1.

```
short_path_lengths = []
counter = {}

# Get the lengths of all shortest paths
for T in nx.shortest_path_length(G_flights):
    for w in T[1].values():
        short_path_lengths.append(w)

# Compute frequency of different shortest path of
     different nodes (airport)
counter = cs.Counter(short_path_lengths)
length_dist = counter
sort_length_dist = dict(sorted(length_dist.items(), key=
     lambda x: x[1], reverse=True))
```

Listing 2: Compute shortest path lengths

Computing the shortest path lengths for each node (i.e. airport) allowed us to visualize the distribution of the shortest path lengths for the graph. Figure 5 illustrates the frequency of each shortest path before the pandemic. Our results show that most nodes' shortest path length was approximately 2.

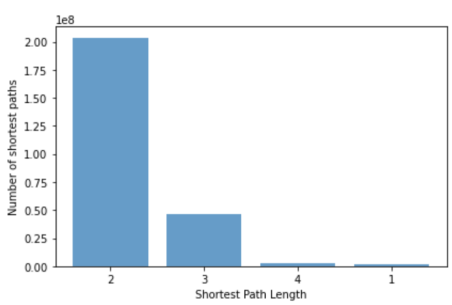


Figure 3: Shortest path lengths distribution before COVID-19.

In contrast, we observed that for the period between February 2020 and September 2020 the distribution of shortest path lengths is slightly different. As is expected, the shortest path lengths increased in size, with the majority of nodes having a shortest path length of 3. For this period, there was also a noticeable increase in the number of nodes with a shortest path length of 4 or 5, as compared to the time period prior to COVID-19.

The plot shown in Figure 4 shows the number of shortest path during COVID-19.

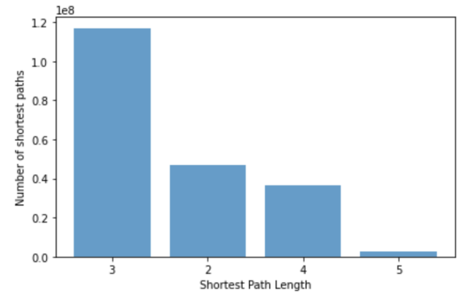


Figure 4: Shortest path lengths distribution during COVID-19.

The *average shortest path lengths* of the giant components for the graphs representing both time periods were calculated. Before the pandemic the average shortest path length was **2.9084392529078866** and during the pandemic it stands at **2.968157191459244**.

These results illustrate a structural change of the global flight network, as airport closures caused the average shortest path length of the flight network to increase by $\tilde{0}$.0597 ($\tilde{3}$%)

6.3 Graph Diameter

The diameter of a graph measures greatest distance between a node *n* and any other node in the graph. In the context of our analysis, the distance between two nodes is the number of edges between a shortest path connecting the two nodes. We calculated the diameter of the graphs representing the time periods before and during the pandemic. The obtained values were **12** and **8** respectively.

Our results indicate that the diameter of the graph representing the time during the pandemic, that is, between February 2020 and September 2020, is lower. One plausible explanation of this structural change is that some airports likely had no incoming or outgoing flights during that time frame, especially airports located in remote locations. More precisely, some edges leading up to the node that was previously reachable with a shortest path distance of 12 may have been removed-were disconnected-from the graph and its giant component.

6.4 Link Analysis

In the following sections we evaluate the results of two link analysis algorithms to answer our research questions related to the importance of airlines and airports, and how their rank or importance may have been affected by COVID-19. More specifically. we ran the PageRank algorithm to determine the importance of each airport (i.e. node) on the flight network as described in the research paper by Xu, Kecheng, et al. [2]. Furthermore, we computed hub and authority scores for each node using the HITS algorithm.

6.5 HITS algorithm - Airports as Hubs

HITS is a link analysis algorithm that was originally developed to rate web pages as hubs and authorities. A Hub score estimates each node's value based on outgoing links/edges. We observed that even though the hub and authority scores differed between nodes, each node had the same hub and authority value. This observation was attributed to the fact that we are using an undirected graph in our analysis which is converted to a directed graph where origin -> destination represents the direction.

Our results obtained from running the HITS algorithm on the flight graphs' giant component are illustrated in Table 1 and Table 2. From the results we obtained, we observed that the Hub values for each node decreased and the order of the top 10 hubs was altered during the time period of COVID-19.

Airport	Hub score	Airport	Hub score
KTEB	0.0017761036426370167	KTEB	0.0016240522787128236
KIAD	0.001660560578816895	KIAD	0.001595467091558386
KMDW	0.0015815128241332549	KPDK	0.0015161571057972045
KHPN	0.0015368896524886933	KBNA	0.0014787884600325916
KPHL	0.0014871217972139581	KHPN	0.0014769255251306027
KORD	0.001473988692681144	KCLT	0.0014717875531392902
KDAL	0.0014527021001370078	KPWK	0.0014637917200364754
KRDU	0.00144634616806044	KDAL	0.001461364577000406
KPDK	0.0014457328069076	KMDW	0.0014338943761651774
KPWK	0.0013950609127753742	KPHL	0.0014226995483519868

Table 1: Top 10 Hub scores before COVID-19

Table 2: Top 10 Hub scores during COVID-19

6.6 PageRank of Airports

PageRank is a popular link analysis algorithm originally developed by Google founders Larry Page and Sergey Brin at Stanford University in 1996 as part of a research project. More generally, PageRank computes a ranking of the nodes in the graph G based on the structure of the incoming links. In our graph, PageRank computes a ranking of airports based on incoming flights.

Airport	PageRank	Airport	PageRank
KORD	0.0020130429307540967	KORD	0.0015859670886150124
KTEB	0.0019138877803737235	KDFW	0.0015216828771316135
KDFW	0.0018278343498018866	KAPA	0.001441626211846016
KATL	0.001725427214503263	KDAL	0.0013749125260387953
KIAD	0.0016889649068276206	KATL	0.0013543625205623775
KDAL	0.0015701779912363574	KIAD	0.001352183748601223
KMDW	0.0015436844171690862	KADS	0.0013481213428859064
KAPA	0.0015009025351049698	KTEB	0.001340040291658249
KPWK	0.0014876191406363632	KSUS	0.0013145974710511582
KHPN	0.0014702998677312564	KCLT	0.0013018947671446428

Table 3: Top 10 PageRank scores before COVID-19

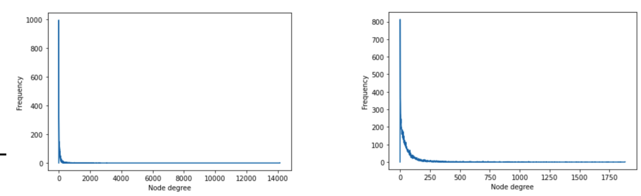
Table 4: Top 10 PageRank during COVID-19

Results show a decrease in the PageRank scores for the top 10 airports for the time period during COVID-19. We have attributed the change in PageRank scores to a significant lower number of flights for this time period.

6.7 Degree Distribution

The degree distribution is the probability distribution of all graph degrees over the entire network. Part of our problem statement is to determine whether the global flight network is scale-free. Understanding the degree distribution of the network allowed us to determine whether this particular network displays a specific degree distribution, such as a Power-law distribution, as many other real-world networks do. In addition, comparing node degrees prior to, and during the pandemic helped identify changes in the number of inbound or outbound flights at each airport since the enactment of global travel restrictions.

During the pandemic, due to the closure of many airports around the world, the node degrees and their frequency changed as illustrated in Figure 6. However, the network remained scale-free and produces a power-law distribution.



(a) Degree distribution before COVID-19.

(b) Degree distribution during COVID-19.

Figure 5: Degree distribution comparison

6.8 Node Degree

In order to know the number of flights at certain airport, the degrees of the nodes (airports) were calculated and then sorted in descending order to know the top 30 busiest airports around the world. Figures 9 and 10 illustrate the number of flights for each airport represented by its unique four letter ICAO code.

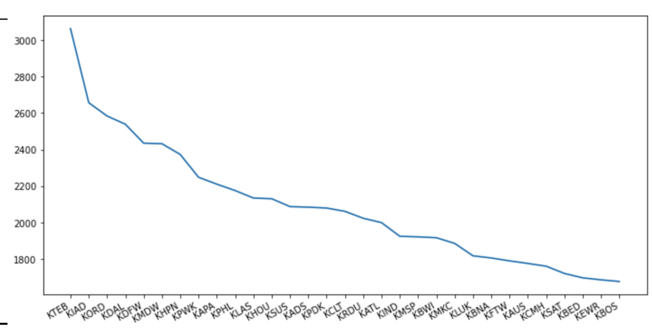


Figure 6: Number of flights for top 30 airports before COVID-19

As shown in Figure 9, we observed that the airport with highest number of inbound and outbound flights before COVID-19 was Teterboro Airport (KTEB), then Dulles International Airport (KIAD), followed by O'Hare International Airport (KORD), Dallas Love Field Airport (KDAL), Dallas/Fort Worth International Airport (KDAL), and the lowest airports in term of number of flight of the top 30 were: Laurence G. Hanscom Field (KBED), Newark Liberty International Airport (KEWR) and Boston Logan International Airport (KBOS).

On the other hand, when the data was plotted for the airports with highest number of inbound and outbound flights during the pandemic, it was noticed that the busiest airports were the same as before COVID-19, but with lower number of edges (total number of inbound and outbound flights).

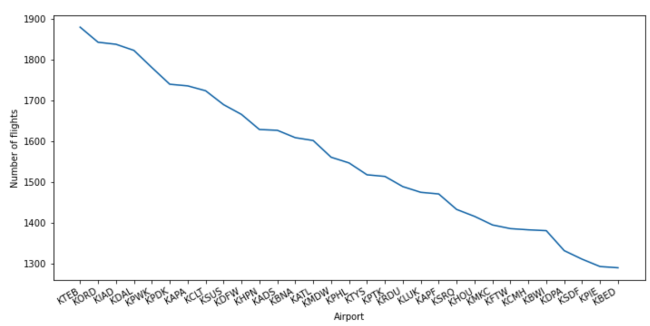


Figure 7: Number of flights for top 30 airports during COVID-19

6.9 Clustering Coefficient

Clustering Coefficient in graph theory is a measure of the degree to which nodes in a graph tend to cluster together. In real-world networks, nodes tend to create tightly knit groups characterized by a relatively high density of ties. In our case, which is the commercial flight network, there are some airports in the network which form high density connection between each others, which means there is a lot of flights between them. Figure 8 and 9 illustrates the Clustering Coefficient of certain airports as following:

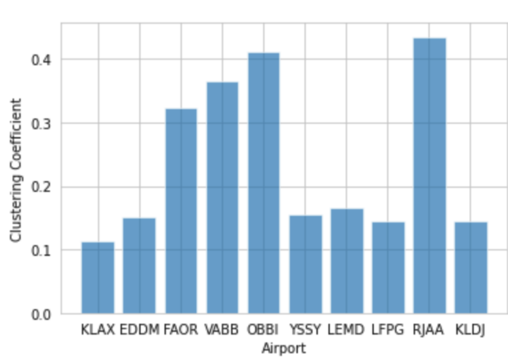


Figure 8: Clustering Coefficient for certain airports before COVID-19

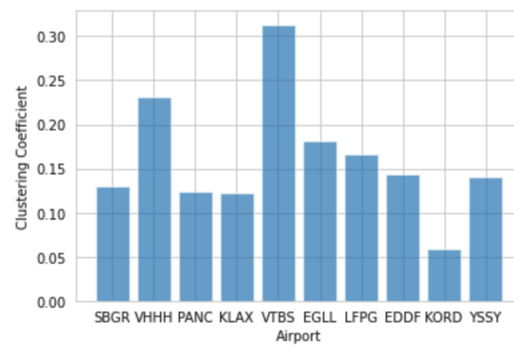


Figure 9: Clustering Coefficient for certain airports during COVID-19

As noticed from the previous two figures, Narita International Airport (RJAA) in Japan had the higest clustering coefficient score (more than 4) before COVID-19, on the other hand, we can notice that Suvarnabhumi Airport (VTBS) in Thailand had the highest clustering coefficient score during COVID-19.

To determine the Clustering Coefficient we used the built-in function from NetworkX Python library (clustering).

```
clusters_of_graph = nx.clustering(flight_graph)
plt.ylabel("Clustering Coefficient")
plt.xlabel("Airport")
x_label_data = list(clusters_of_graph.keys())[:10]
y_label_data = list(clusters_of_graph.values())[:10]
#x_label_data = np.arange(len(x_label_data))
plt.bar(x_label_data, y_label_data, align="center", alpha
    =0.6)
```

Listing 3: Compute clustering coefficient

6.10 Link Prediction

Link Prediction is an important approach to what is called (Link Mining). Link Prediction is used to predict whether there will be links (routes) between two nodes (airports) based on the observed existing link information.

Part of our motivation in this project was to determine whether we can predict changes in the network due to the effects of the pandemic. More specifically, can we devise a simple algorithm to predict link removal from the network, based on data prior to COVID-19 restrictions and compare this to what we see in the real-world network for the time period affected by the pandemic?

The algorithm is outlined in Listing 4 and Listing 5. It can be summarized in five main steps:

Given a graph G containing flight data:

- Compute edge betweenness values for graph G (data prior to COVID-19)
- Sort edges by betweenness value in ascending order
- Select K edges
- Search for edge Ek in graph G (data during COVID-19)
- If edge Ek is not present, we correctly predicted the flight route cancellation.

```
from networkx.algorithms.centrality import
    edge_betweenness_centrality
```

```python

# Calculate edge betweenness for the GCC of the graph
y = edge_betweenness_centrality(giant_one, k=1000)

r = cs.Counter(y)
r_copy = r.copy()

# Remove edges that are using waypoint nodes
for i in r.keys():
    u = i[0]
    v = i[1]
    n = r[i]
    # Remove WAYPOINTS from the data.
    # Nodes with NUMBERS are not airports, they are Navigation waypoints.
    if (not u.isalpha() or not v.isalpha()) or n == 0.0:
        #print("{} {} {} bad".format(u,v,n))
        r_copy.pop(i)

# Sort and select 10,20 and 100 edges with
# lowest betweenness scores
predicted_20 = r_copy.most_common()[-20:]
predicted_10 = r_copy.most_common()[-10:]
predicted_100 = r_copy.most_common()[-100:]
```

Listing 4: Link prediction algorithm

The method in Listing 5 calculates the Precision at K values using a set of predicted removal edges. The set of predicted edges are obtained from running the prediction algorithm on a graph built using data for the period prior to COVID-19. The precision@k is then calculated based on the existence of those predicted edges in the graph built using data for the time period during COVID-19.

```python
def precision(predicted, k):
    total_accurate = 0
    for i in predicted:
        u = i[0][0]
        v = i[0][1]
        if not giant_one.has_edge(u,v):
            total_accurate += 1

    precision = total_accurate / k
    print("Precision@20 = {}".format(precision))
```

Listing 5: prediction@k

Our experiments with k=10, k=20 and k=100, where k is the number of predicted edges selected for removal were:

Precision@20 = 0.5

Precision@20 = 0.45

Precision@20 = 0.41

The accuracy of the results are restricted by the value of k (the set of edges with lowest betweenness values that are selected to compute accuracy) and the k value in the edge_betweenness() NetworkX method that determines the number of edges used to calculate centrality values. In future work, our algorithm could be improved by experimenting with different values for k, and improving the efficiency of the edge betweenness calculation method from NetworkX. This method could be more efficient if it utilized multiple threads or processes to calculate centrality values by chunking the data and performing computations in parallel. A solution that uses a map-reduce paradigm may be a better approach when implementing this method.

7 CONCLUSION

First, we considered the problem of analyzing a real-world network in two different time periods to identify structural changes caused by a real-world event such as the COVID-19 pandemic. We proposed a methodology for building a graph representing the global commercial flight network based on open and free data provided by the OpenSky Network. We identified numerous changes to the network structure caused by a reduction in the number of flights across the network. More specifically, for the period between February 2020 and September 2020, the average shortest path length of the network increased by approximately 3%, while the diameter of the graph was reduced from 12 to 8, a 25% reduction. Increasing the average shortest path would increase the flight duration to reach the flight destination. As shown in fig 4, we observed that the most frequent shortest path was 3, which is higher than the most frequent shortest path before COVID-19, which was 2. We observed a decrease in the diameter of the flight network during COVID-19. We attribute this change to the fact that edges were removed between different airports due to flight restrictions.

In addition, the degree distribution of the flight network varied slightly before COVID-19 and during COVID-19. As shown in fig 5, both graphs have scale-free degree distribution, but the node degree values were lower, as was their frequency. This proved that the number of inbound and outbound flights for different airports decreased. Figures 6 and 7 also support this observation. We also observed that the number of flights for different airline companies was significantly reduced during COVID-19.

Second, we ran two link analysis algorithms, PageRank and HITS, to identify important airports. We observe that the scores produced by the algorithms are lower for the time frame of the pandemic and 5 out of the top 10 airports that were scored by PageRank were different. This might prove that some airports were affected a lot by the pandemic. For example, KTEB was scored the second most important airport before the pandemic, and it was scored the 8th top airport during the pandemic. By contrast, KAPA observed a higher PageRank score during COVID-19 and moved from 9th place to 3rd place.

Finally, our link removal prediction algorithm that utilizes low edge betweenness values as the main indicator for removal shows promising results but further work is required to optimize the algorithm and better estimate its accuracy.

8 REFERENCES

(1) “Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis,” COVID-19 - All Documents. [Online]. Available: https://www.icao.int/sustainability/Documents/COVID-19/ICAO_Coronavirus_Econ_Impact.pdf [Accessed: 07-Oct-2020]

(2) K. Xu, T. Long, S. Qiao, Y. Zheng and N. Han, "A Global Flight Networks Analysis Approach Using Markov Clustering and PageRank," 2017 IEEE International Conference on Big Knowledge (ICBK), Hefei, 2017, pp. 96-102, doi: 10.1109/ICBK.2017.42.

(3) T. Suzumura, H. Kanezashi, M. Dholakia, E. Ishii, S. A. Napagao, R. Pérez-Arnal, and D. Garcia-Gasulla, “The Impact of COVID-19 on Flight Networks,” arXiv.org, 05-Jun-2020.

[Online]. Available: https://arxiv.org/abs/2006.02950v2 [Accessed: 07-Oct-2020].

(4) X. Olive, M. Strohmeier, and J. Lübbe, “Crowdsourced air traffic data from The OpenSky Network 2020,” Zenodo, 17-Sep-2020. [Online]. Available: https://zenodo.org/record/4034518 [Accessed: 07-Oct-2020]

(5) “Industry Losses to Top $84 Billion in 2020,” IATA. [Online]. Available: https://www.iata.org/en/pressroom/pr/2020-06-09-01/. [Accessed: 06-Nov-2020].

(6) “After April Passenger Demand Trough, First Signals of Uptick,” IATA. [Online]. Available: https://www.iata.org/en/pressroom/pr/2020-06-03-01/. [Accessed: 06-Nov-2020].