Universität Zürich[UZH]

Blockchain & Distributed Ledger Technologies

Final Report: How a pandemic spreads over the flight system

David Diener (19-733-179), Rafael Dubach (18-755-140), Layla Husselman (17-733-130), Kevin Kindler (15-922-529)

Network Science

Faculty of Business, Economics and Informatics

December 2022

ABSTRACT

Pandemics are known to spread rapidly over the world if no measures are taken to control the spreading. In this final report, we analyze the flight network through centrality measures to detect nodes of interest. With the help of a spreading simulator, we then analyze the behavior of said nodes in a pandemic setting. To assess the behavior of the network without the node of interest, we ran the simulation without these nodes. Overall, our results show that the removal of nodes of interest is an effective strategy to reduce the spread of a pandemic. By targeting the most connected parts of the network, we can significantly reduce the reproduction rate of the disease and limit its spread.

Click here to open the github repository.

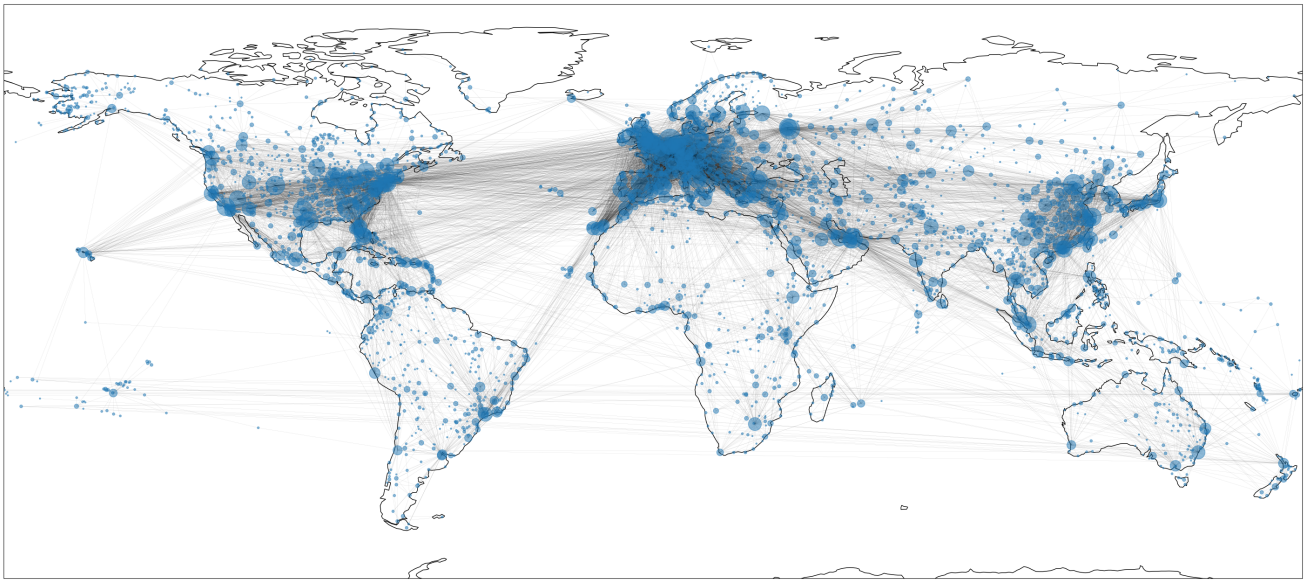

Fig. 1: Heatmap of the world picturing all airports with their degree as a bubble. Edges between airports displaying the routes of airlines.

1 INTRODUCTION

The focus of this report is on the study of the spread of pandemics over the flight network. Pandemics are widespread occurrences of a particular disease in a population, and the spread of pandemics is a topic of significant concern to public health officials and researchers. As globalization continues to increase and travel becomes more affordable, it is increasingly important to understand the mechanisms and dynamics of disease spread in order to design effective intervention strategies. In this report, we present a study on the spread of pandemic over the flight network using network science techniques. We begin by introducing the dataset used in the study, which consists of a network of airports and flights routes. Also, we provide connections to relevant literature regarding disease spreading in networks. We then present the results of our analysis, which includes degree, centrality, and PageRank measures to identify key nodes in the network that are critical for controlling the spread of a disease. Additionally, we introduce a spreading simulator to illustrate how a disease can spread from these critical nodes. These techniques enable us to answer the following research question:

RQ: Which nodes (airports) would be critical to control the spreading of the pandemic?

Our findings provide insights into the mechanisms and dynamics of disease spread in the flight network and can inform the design of effective intervention strategies and improved preparedness for future outbreaks.

1.1 Data

The data we used in our model comes from DAFIF (Digital Aeronautical Flight Information File) and represents flight information from all over the world. The data-sets were made available through openFlights (*Airport, airline and route data* 2017). An entry in the airport data-set may look as described in Figure 2. It contains key information regarding each airport e.g: an airport ID, the airport's name, country, and geographical information (i.e. latitude and longitude).

Example Node	
Name	Description
Airport ID	Unique OpenFlights identifier for this airport.
Name	Name of airport. May or may not contain the City name.
City	Main city served by airport. May be spelled differently from Name.
Country	Country or territory where airport is located. See Countries to cross-reference to ISO 3166-1 codes.
IATA	3-letter IATA code. Null if not assigned/unknown.
ICAO	4-letter ICAO code. Null if not assigned.
Latitude	Decimal degrees, usually to six significant digits. Negative is South, positive is North.
Longitude	Decimal degrees, usually to six significant digits. Negative is West, positive is East.
Altitude	In feet.
Timezone	Hours offset from UTC. Fractional hours are expressed as decimals, eg. India is 5.5.
DST	Daylight savings time. One of E (Europe), A (US/Canada), S (South America), O (Australia), Z (New Zealand), N (None) or U (Unknown). See also: Help: Time
Tz database time zone	Timezone in "tz" (Olson) format, eg. "America/Los Angeles".
Type	Type of the airport. Value "airport" for air terminals, "station" for train stations, "port" for ferry terminals and "unknown" if not known. In airports.csv, only type=airport is included.
Source	Source of this data. "OurAirports" for data sourced from OurAirports, "Legacy" for old data not matched to OurAirports (mostly DAFIF), "User" for unverified user contributions. In airports.csv, only source=OurAirports is included.

Fig. 2: Sample entry of the airport data-set

Correspondingly an edge contains the information found in Figure 3 below. Edges have a direction (i, j) which indicates a regularly occurring commercial flight by a particular airline from airport i to airport j. It is possible that multiple edges exist between a pair of airports if multiple airlines offer a flight on the same route, or if one airline offers multiple such flights each day, but with different flight numbers (i.e. counts as a separate route). Kunegis 2017 stresses the fact that a flight in this data-set is defined as a normally regularly occurring flight identified by its flight number (e.g., AF331), not individual flights (e.g., AF331 on June 14 2015).

Example Edge	
Key	Value
source	0
target	2
distance	124.520121
airline	CG
airline_code	1308
codeshare	0
equipment	DH8 DHT
stops	0

Fig. 3: Sample entry of the routes data set

For our analysis, the airports (nodes) and routes (directed edges) between two airports were relevant. To form the graph $G(V, E, w)$ we first added all nodes to the graph, and then iterated through the edges in the routes data-set and added corresponding edges manually (see Listing 1).

```
for a,b in edges.iterrows():
    if G.has_edge(b['#source'], b['target']):
        G[b['#source']][b['target']]['weight'] += 1
    else:
        G.add_edge(b['#source'], b['target'], weight = 1)
```

Listing 1: Adding edges to the graphs

As seen in the listing above, weights are also manually added. The default weight is 1 and it will be increased by the same amount if another airline (or the same with another flight number) also offers this route. Through this process, the routes with several flights get a bigger emphasis.

2 THEORY

The structure of complex networks (e.g. the flight network) has long been understood to play a vital role in containing the spreading of pandemics. Since these networks are a crucial part of our society, it is important to understand their inner workings, to learn to handle appropriately in such events.
Harper and Tee 2020 have analyzed how route planning and capacity management affects the spreading of diseases. Through weighted vertex entropy or degree scoring of nodes (airports) they were able to reduce the reproduction rate of diseases way better in contrast to a random airport and route closure scenario. In the paper, they also touch on the fact that the friendship paradox (i.e. the friendship paradox is the observation that friends of individuals tend to have more friends or be more popular than the individuals themselves (Pal et al. 2019)) holds in an airline network since the degree distribution is not homogeneous.

Bielecki et al. 2021 analyzed air travel and COVID-19 prevention in the pandemic from a broader (i.e. policies, and prevention) perspective. They found that disease transmission in aircraft are to be considered very low, as aircrafts are equipped with high-efficiency filtering, temperature and humidity monitoring, which all help to mitigate the spreading. Nevertheless, over 27 million travelers, confirmed in-flight cases had been published. Their conclusion showed that travel restrictions only have a limited effect in containing the spread of a pandemic. The restrictions' impact is also highly dependent on several factors, such as the extent and timing of the

restrictions, the pandemic size, the virus' reproduction number, and travel patterns. The authors suggest that the best way to contain the spread is to facilitate saliva testing on arrival, to allocate quarantine accordingly.

Zhou et al. 2021 have analyzed the impact of network topology on air transportation to pandemics and have found that it is accurate that having more international airports can increase a country's international connectivity. However, the proportion of cities with international airports is the most dominant topological metric for determining a country's international robustness towards pandemics. Further research (Zhang et al. 2017) has also covered that there are many other factors that can affect a country's international connectivity and robustness, such as the quality and capacity of its airports, the number and strength of its international partnerships and agreements, the ease of obtaining visas, and the country's overall economic and political stability. Additionally, the proximity of a country to other major international hubs and its transportation infrastructure can also play a significant role in its international connectivity. In short, the international robustness of a country is determined by a complex interplay of various internal and external factors, and the proportion of cities with international airports is the strongest but by far not the only factor that has an effect on it.

3 METHODS AND RESULTS

Here, we present the results of our analysis and discuss the implications of our findings. For the directed graph under consideration, the Pearson correlation coefficient of degree between pairs of linked nodes is -0.01918. This indicates that the assortativity coefficient of the network lies between -1 and 1, indicating that the network is neither highly assortative nor highly disassortative but exhibits a moderate level of assortativity which concludes that the network has good assortative mixing patterns. This suggests that some nodes in the network are well-connected, while others are not. In the following Figure 4, we list some other global properties calculated from the constructed flight network.

Directed graph measures	
Density	0.003573980238572647
Average clustering	0.4782853255477226
Number of nodes	3214
Number of edges	36907
Average degree	6.7238493723849375
Max weight	20

Fig. 4: Calculated measures for directed graph

3.1 Degrees

In a graph, the in-degree of a node is the number of incoming edges to that node, while the out-degree of a node is the number of outgoing edges from that node. In-degree and out-degree are commonly used to describe the connectivity of nodes in a network. For example, in a network of airports and routes, the in-degree of an airport would be the number of routes incoming at that airport, while the out-degree of the airport would be the number of routes departing from that airport. In a graph, the in-degree and out-degree of a node can be represented visually by the number of arrows coming into or going out of the node. So it is logically expected that the in-degree and out-degree of an airport are similar, as an airline that arrives at an airport with a flight will typically continue flying with the same aircraft and therefore also have a route that departs from the airport. Therefore, it is unsurprising that the 10 airports with the most arrival routes are also the 10 airports with the most departure routes. These airports are Frankfurt am Main, Charles de Gaulle International Airport, Amsterdam Airport Schiphol, Atatürk International Airport, Hartsfield Jackson Atlanta International Airport, Chicago O'Hare International Airport, Beijing Capital International Airport, Munich Airport, Dallas Fort Worth International Airport and Domodedovo International Airport. The degree distribution of in and out degrees of the whole network is displayed below.

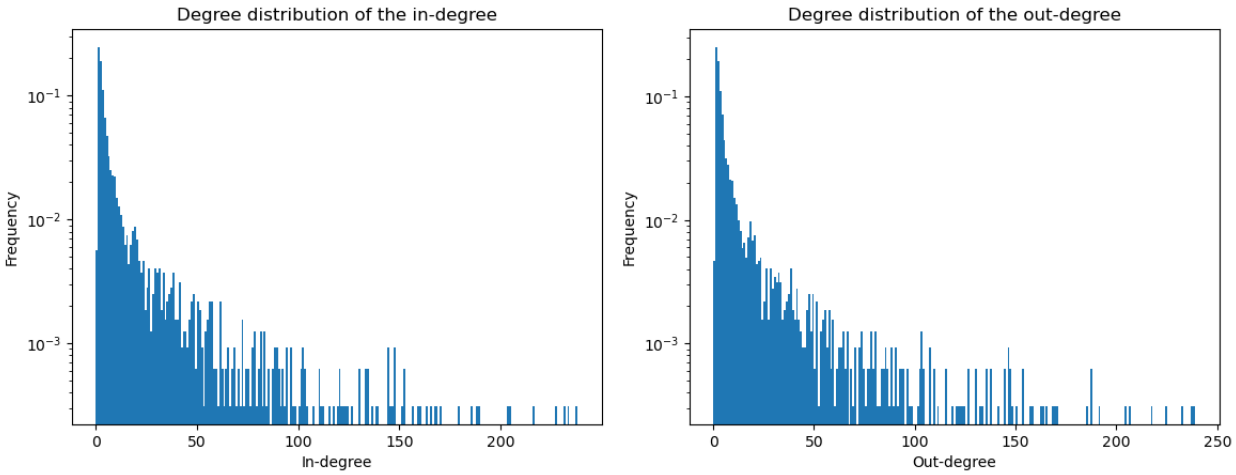

Fig. 5: Degree measures - results

3.2 Centrality

The four centrality measures in a graph are degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Degree centrality is a measure of the number of connections a node has in a network. In other words, it is a measure of the node's in-degree and out-degree. Nodes with a high degree centrality are often considered important or central in the network, as they are connected to many other nodes. Closeness centrality is a measure of how close a node is to all other nodes in the network. Nodes with a high closeness centrality are able to reach other nodes quickly and easily, making them important for the flow of information or resources in the network. Betweenness centrality is a measure of how often a node lies on the shortest path between two other nodes in the network. Nodes with a high betweenness centrality are often considered important, as they often act as intermediaries between other nodes in the network. Eigenvector centrality is a measure of the importance of a node in a network based on the importance of its neighbours. Nodes with a high eigenvector centrality are connected to other important nodes, making them important in the network as well. Overall, these centrality measures can be used to identify important nodes in a network and to understand the flow of information or resources in the network. To identify the most important airports in the dataset, we calculated all four centrality measures (degree, closeness, betweenness, and eigenvector centrality) for all nodes in the network. Our analysis indicates that the airports with the highest centrality measures across all four measures are Frankfurt am Main Airport, Charles de Gaulle International Airport, Amsterdam Airport Schiphol, and Atatürk International Airport. These airports also appear in the list of 10 highest in/out-degree distributions mentioned earlier. Overall, our

analysis suggests that the airports with the highest centrality measures in the network are also the busiest and most important airports in the dataset. These airports serve as key hubs for both domestic and international flights, connecting passengers to a wide range of destinations and playing a crucial role in the global aviation industry. The results of our analysis are displayed below.

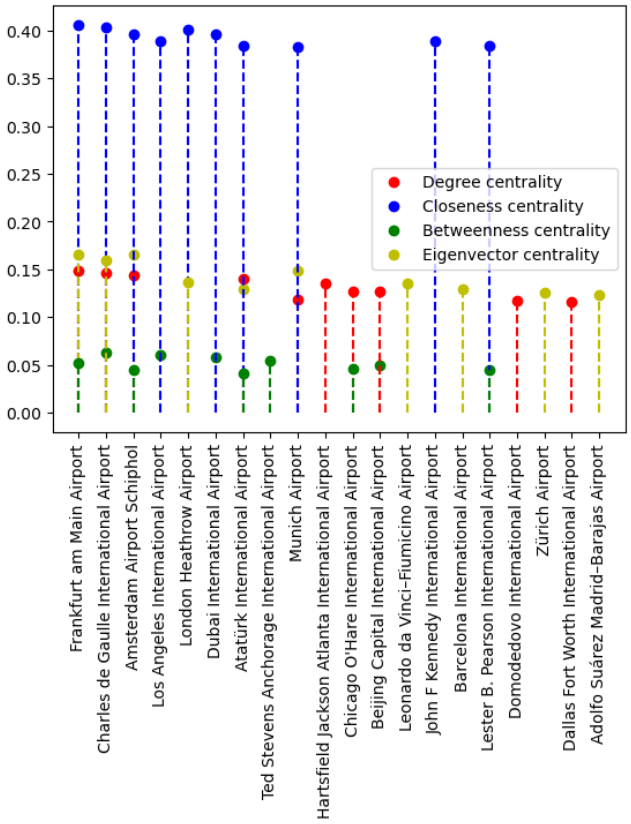

Fig. 6: Centrality measures - results

3.3 PageRank

In addition to the centrality measures, we calculated the page ranks of the airports in the network using the PageRank algorithm developed by Larry Page of Google. This algorithm is a method for determining the significance or relevance of a network node based on the quantity and quality of links pointing to that node. Although the method was originally used to measure the significance of websites, there are numerous research papers in which the algorithm was used to identify key nodes in a network spreading like Bhattacharya et al. 2021, which identified hotspots of COVID-19 in the Indian subcontinent through the page rank algorithm. According to Brin and Page n.d., the page rank parameter alpha is a damping factor that can be set between 0 and 1 and is typically set to 0.85. It is the probability at each page that a person surfing the web ”random surfer” will get bored and requests another random page (continue surfing). As a consequence in terms of spreading, the damping factor alpha is the probability, at any step, that the disease will continue spreading and 1-alpha is the probability, at any step, that the disease will stop spreading. For 11 alpha values between 0 and 1 the 10 airports with the highest page ranks were extracted and plotted as seen in figure 7. It can be seen that three airports consistently have high page ranks for all values of alpha between 0 and 1: Hartsfield Jackson Atlanta International Airport, Dallas Fort Worth International Airport, and Chicago O’ Hare International Airport. These airports are consistently ranked among the top 10 airports with the highest page rank, indicating that they play a significant role in the network. Furthermore, we find that the four airports previously identified as the most important based on centrality measures and in-/out-degree distribution also have page ranks among the 10 highest for certain alphas. Specifically, Frankfurt am Main Airport for alpha between 0.9 and 0.99, Charles de Gaulle International Airport for alpha between 0.7 and 0.99, Amsterdam Airport Schiphol for alpha of 0.99, and Atatürk International Airport

for alpha between 0.4 and 0.6. This indicates that at least the first three of these airports are consistently ranked as being among the most important in the network when dealt with a very contagious disease, regardless of whether measured with centrality or the page rank. Nevertheless, the fourth airport Atatürk International Airport has very good results in the other measurements and does also occur in the top 20 with respect to the page rank (although for less contagious diseases) therefore, this airport was added to our favorites in regard to importance.
Overall, the analysis of page ranks provides valuable additional insight into the structure of the airport network. By examining the page ranks of the airports, we can identify key hubs and assess the relative importance of different airports in the network.

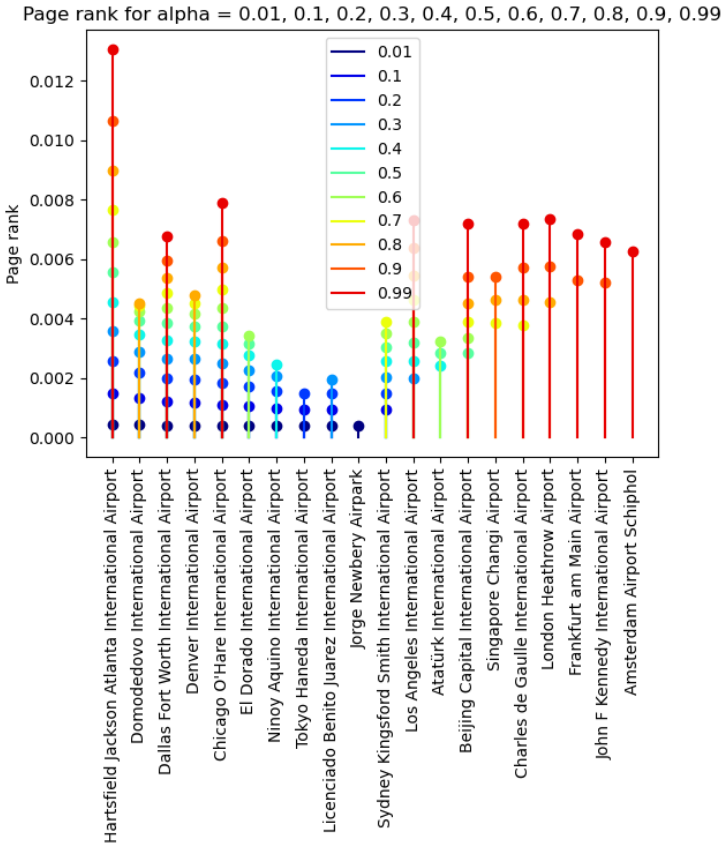

Fig. 7: Page rank measures - results

3.4 Spreading Simulator

Measuring the rate at which a disease spreads can be challenging due to the complex interactions and variables involved. One method that is often used is to employ a proxy measure, such as the number of iterations needed until all nodes have been reached and the spreading of the disease stops. We modified a Breath First Search (BFS) algorithm to simulate the spreading across the network. This spreading simulator takes the starting airport from which the disease starts spreading and a spreading factor between 0 and 1 as input. If the spreading factor is set to 1 (100%), the disease will spread to all neighbouring airports and reflects a high level of contagion. If it is set to 0 non of the neighbouring airports will be contaminated and if it is between 0 and 1 the disease is randomly distributed among the percentage of neighbors. The importance of the input airport from which the spreading starts can be measured of the amount of levels (depth) which is needed until the spreading stops (disease has spread to all possible airports). The fewer levels (smaller depth) are needed to reach all possible airports, the faster the disease spreads and the more important is the starting airport. To illustrate this in a graph, on every level the airport nodes are assigned a different color (see figure 8). This approach allows us to gain a better understanding of the dynamics of disease spread and to identify potential interventions that may be effective in reducing the spread of the disease.

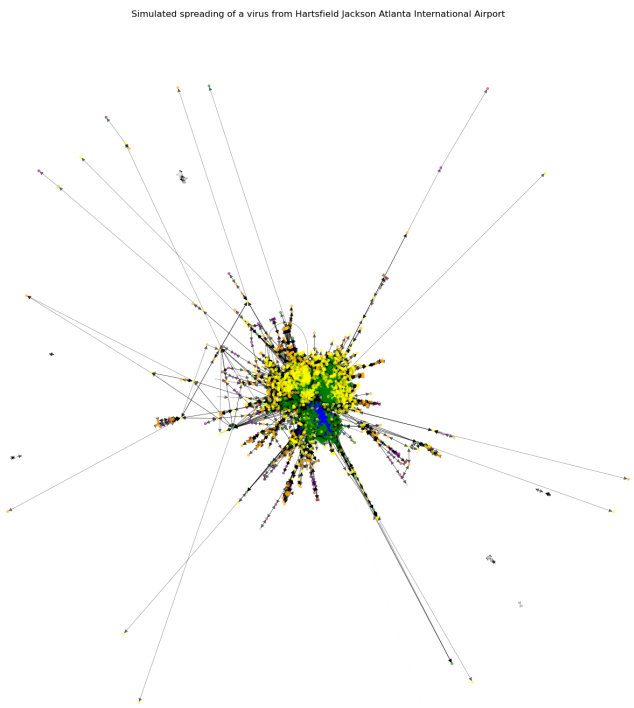

Fig. 8: Spreading of the airport Hartsfield Jackson Atlanta International Airport with the highest page rank

The spreading simulator allowed us to interpret the results form sections 3.1, 3.2 and 3.3 in a different way. When running the simulator for each airport once with the original network (blue) and once for the network from which the most important airports identified above were removed (red), it is possible to find out what impact the removement of the airports has on the spreading. The result is plotted in Figure 9. It is noticed that only the airports are taken into account which can spread to at least 98 percent of all the airports in the network. In this way small isolated airport groups have been removed from the big main group. It can clearly be seen that by removing only the 4 most important airports of the main network, namely the network without the isolated airport groups (= 3166 airports), the spreading simulator reaches a higher depth for a lot of the airports. By example, there are not anymore 70 airports from which the disease can spread to 98 % of the airports in only 7 iterations but 36 starting airports. Also the disease can not spread anymore to 3166 airports but only to 3145.

Fig. 9: Spreading within the network with and without the 4 most important nodes

4 DISCUSSION

The purpose of our study was to examine the aviation network in relation to the transmission of pandemics, with a particular focus on airports (the nodes of our network) and the significance of certain airports for the spread of the disease. According to the results of the in-degree and out-degree analysis, as well as the centrality measurements and the PageRank, there are four airports that warrant close examination: Frankfurt am Main Airport, Charles de Gaulle International Airport, Amsterdam Airport Schiphol and Atatürk International Airport. These four airports are significant for a pandemic's spread:

- According to the centrality metrics, these airports are network hubs since they are connected to numerous other airports (which makes the spreading faster).
- They are also close to a large number of other airports within the network because of high closeness centrality, which again expedites the spreading of the disease by reducing flight durations.
- These four airports act as intermediaries between other nodes (due to high betweenness centrality scores), which makes it more likely for an airplane to land at one of these four airports and continues to fly to the next airport (However, this is not a very good measure for our network since the passengers and crew most likely change between the flights).
- The high eigenvector centrality characterizes the connectivity of these four important airports to other airports that are also important.
- The PageRank values of three of this four airports are under the 20 highest page ranks for high alphas (very contagious disease) and the fourth airport has high page ranks under the 20 highest page ranks for moderate alphas (little contagious disease).

In addition, the significance of these four airports was investigated by removing them from the data set and determining whether additional iterations are required to infect all airports. Indeed, this was the case. Through removing these airports, many airports, which needed 7 iterations now needed 8 or more. We can therefore conclude, that these four airports are crucial for controlling the spread of a pandemic. The spreading simulator highlights this discovery. It was run for all of the four important airports, as well as for some of the less important airports for comparison. If the infection chance was set to 100%, each of the four most important airports would require precisely seven iterations. The randomly selected comparator airport (El Tari Airport) required two more repetitions, namely nine. Although the majority of airports require more than seven cycles, certain airports only require seven iterations (Figure 9).
In the real world, the infection of a pandemic is never 100%. It depends on the illness and many other external factors. When the infection probability was set to an extremely low value, 1%, then the four most significant airports, if set as starting airport, again infected more airports than five randomly selected airports (which infected

no or very few other airport). Due to the low infection probability, the iterations of the randomly chosen airports were quite low, the pandemic stopped spreading quite fast. However, for the four main airports, there were on average more than ten iterations before the pandemic died out. This indicates that these four airports are indeed more important to control in the event of a pandemic, as the disease spreads more rapidly through these airports.

However, this study is a highly simplified model of the real world. Since there was no publicly available data source containing all flights, we utilized the data set provided by OpenFlights (*Airport, airline and route data* 2017). This data set provides a lot of important information about the airports and their connections to other airports. To get a better understanding of the transmission of a pandemic, statistics of individual flights, ideally including the number of passengers that attended that flight are needed. With this information, the likelihood of disease's transmission might be determined more accurately.
In this study, it was assumed, that when there are several routes between two airports, there are also more flights. The number of routes between two airports was therefore defined as the weight of the edges (which correspond to the flight routes). However, in the real world there may be multiple flights within one route, which would return a different result. Moreover, only passenger flight routes were considered within our dataset (*Airport, airline and route data* 2017). However, it is also possible to spread the disease via cargo flights, by having infected crew members. In addition, it was assumed, that the destination airport either got infected or not. In the real world, it would also be possible to bypass an airport and infect the next (perhaps because a sick passenger did not infect anyone on a stopover airport or on the flight to this stopover airport). If an airport does not get infected and has connections to several other airports that are not connected to the main network through other nodes, these also won't get infected. In the real world however, this would be possible since an airport has several flights that can spread the pandemic to these airports. Because of this limitation, we mainly focused on the spreading for probability set to 1 (100%). Within a pandemic it is almost certain that at one point or another, every airport within the aviation network gets infected.
Another limitation of this study is that an airport is either infected or not, but can not recover and get reinfected. This is not a severe limitation as the transmission of a pandemic is relatively rapid and the airports, or the cities in which they are located, mostly only get infected, are therefore not able to recover within the spreading of a pandemic and can pass on the pandemic. Therefore, it is unlikely that a city with an airport would become infected during the spread of a pandemic, recover, and then become infected again after recovering. However, if a city with an airport became contaminated, airplanes departing from this airport do not necessarily include one of the infected individuals. For this, we would also need to know the amount of the city inhabitants to get a probability of having a sick individual on a flight, but this would be out of scope for this study since this would also need a spreading simulator within the city of an airport itself.

5 CONCLUSION

This final report studies the importance of nodes (airports) in controlling the spreading of a pandemic. The in-/out-degree distribution, several centrality measures and the page rank were calculated and used to identify the relevant nodes in the aviation network. As discussed in Chapter 4, four airports were identified as vital for the flight network in a pandemic setting.
As stated in Chapter 4, the results need to be enjoyed with care, since numerous variables come into play, when pinpointing nodes of interest in a pandemic (Chapter 2). Nevertheless, the results convey an emphasis on airports that are highly interconnected (i.e. hubs). With the spreading simulator, we were also able to test what effect the removal of the identified nodes has on the rapidity the disease propagates through the network. The fact, it takes most of the iterations (i.e. starting from one node and spreading through the whole network) one iteration longer, is also an identification of the significance of the extracted nodes. Overall we can conclude that it is possible to identify the most relevant nodes of the air travel network on a high level (i.e. measurements through routes) and analyze their significance in a pandemic.

REFERENCES

Airport, airline and route data (2017) https://openflights.org/data.html, Last visit Novemeber, 2022.

Bhattacharya, Sawon, S. Siva Sathya, and S. Sharmiladevi (2021) "Analyzing the Spread of COVID-19 in India Through PageRank and Diffusion Techniques". In: *Emerging Technologies in Data Mining and Information Security*. Ed. by João Manuel R. S. Tavares, Satyajit Chakrabarti, Abhishek Bhattacharya, and Sujata Ghatak. Singapore: Springer Singapore, pp. 723–733. ISBN: 978-981-15-9774-9.

Bielecki, Michel, Dipti Patel, Jochen Hinkelbein, Matthieu Komorowski, John Kester, Shahul Ebrahim, Alfonso J. Rodriguez-Morales, Ziad A. Memish, and Patricia Schlagenhauf (2021) "Air travel and COVID-19 prevention in the pandemic and peri-pandemic period: A narrative review". In: *Travel Medicine and Infectious Disease* 39, p. 101915. ISSN: 1477-8939. DOI: https://doi.org/10.1016/j.tmaid.2020.101915.

Brin, Sergey and Lawrence Page (n.d.) "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: ()

Harper, Robert and Philip Tee (2020) *Balancing Capacity and Epidemic Spread in the Global Airline Network*.

Kunegis, Jerome (2017) *OpenFlights (Patokallio)*. http://konect.cc/networks/openflights/, Last visit December, 2022.

Pal, Siddharth, Feng Yu, Yitzchak Novick, Ananthram Swami, and Amotz Bar-Noy (2019) "A study on the friendship paradox - quantitative analysis and relationship with assortative mixing". In: *Appl. Netw. Sci.* 4.1, p. 71. DOI: 10.1007/s41109-019-0190-8.

Zhang, Yahua, Anming Zhang, Zhenran Zhu, and Kun Wang (2017) "Connectivity at Chinese airports: The evolution and drivers". In: *Transportation Research Part A: Policy and Practice* 103, pp. 490–508. ISSN: 0965-8564. DOI: https://doi.org/10.1016/j.tra.2017.05.026.

Zhou, Yaoming, Siping Li, Tanmoy Kundu, Xiwen Bai, and Wei Qin (2021) "The Impact of Network Topology on Air Transportation Robustness to Pandemics". In: *IEEE Transactions on Network Science and Engineering* 8.3, pp. 2249–2261. DOI: 10.1109/TNSE.2021.3085818.

AUTHOR CONTRIBUTIONS

All authors conceived and designed the project idea. D.D. performed the literature review and wrote the introduction and the theory. R.D. wrote the methods and results except for the spreading simulator and the page rank, which were written by both, R.D. and K.K. The discussion was written by L.H. All authors discussed and reached the conclusion.

R.D. provided the coding infrastructure. R.D. and K.K performed the data collection and analysis. K.K. defined and plotted the most important airports and analysed them. K.K. and L.H. wrote and analyzed the Spreading simulator. R.D coded the first map whilst L.H. did the visualisation of the second and third map. All authors analysed the data. All authors revised and accepted the final version of this document. The distribution of tasks was fair for everyone.