A Network Topology Analysis of the Airline Industry

Leonore Valentine Guillain; Francesco Pase; Ying Zhuang; Cosmin-Ionut Rusu
Project report for: A Network Tour of Data Science , 2018

***Abstract*—The network of airlines spans the entire globe and enables us to travel from Greenland to Australia with only a few stops in-between. This network is composed of sub networks corresponding to individual airlines. In this project, we analyze these airline networks to gain insight into what the structure of an airlines flight network tells us about the airline. We discover how the topology of airline networks correlates with their price class and geographical location. We also discover that these same topological features in combination with robustness measures strongly correlate with the airlines average delay times. Using this information, we find competitors and give recommendations on how airlines could form strategic partnerships to increase their competitive advantage, by increasing the reach of their network.**

I. Introduction

Much analysis has been done on the structure of the overall network formed by airports and the flights in between. It is a well known example of a scale-free network. The networks formed by individual airlines themselves are also of interest of study but their topology is less well known. Thus we investigate their structure and pose the hypothesis that the structure of an individual airline graph gives insight into the business model of the airline. This analysis is even of more importance as air travel has become ubiquitous in the last decades and competition between airlines is fierce as ever with low-cost airlines establishing themselves. For this reason we specifically analyze whether low-cost airlines have a specific structure, and how it differs from airlines renowned for their services. Additionally, we decide to explore average delay times, which are an important indication of the quality of the airline and relevant for all customers, regardless of them travelling low-cost or first class. With this information in hand, we analyze the competition of airline with a similar business model and network topology. Then, we find ways that airlines can collaborate with each other to beat the competition, again considering airlines with similar business models.

II. Data Collection and Augmentation

We use three datasets collected from Openflights [1] to build the graphs. The routes, containing the list of flights from one airport to another; airports, containing information about individual airports, and airlines, containing information about the airlines. It contains over 10000 airports and 67663 flight routes. The airports dataset was gathered from October 2006 to January 2017 and the routes dataset from October 2006 to June 2014. We gathered a list of low-cost carriers (LCC) [2], and top rated airlines [3], as well as a list of average airline delays [4] to validate the results of the topology analysis. In this project we only consider sufficiently large airlines, which we define as airlines with over 150 flight connections. The reasoning behind this choice is simple: it is harder to quantify the differences in structure for small graphs. When considering individual airlines, we found airports listed in the routes dataset but not in the airlines dataset. They should not be discarded as they often are hubs in the network of an airline. The reason for this inconsistency is that the routes dataset has not been updated as recently as the airport dataset. To remedy this, we augment the list of airports in two ways: firstly, by crawling a Wikipedia page [5] containing most of the relevant data. Secondly, by manually collecting further missing data. Going further we augment the dataset of routes by adding the flight distance between two airports as well as whether the flight is international or not. We can easily infer the latter by checking whether the countries of the airports connected by that flight are the same or not. To get the distance we use the geopy library, which provides us with a distance function, taking two points with longitude and latitude as an input and returning the distance between two points. Then, the graph of an airline is constructed as follows: each node represents an airport serviced by that airline, each edge representing a flight between the two airports. The graph is unweighted as we only have one flight per airline and per route.

III. Methods

A. Preliminary Analysis

We first perform a basic analysis of the structure of airline graphs. Just like the network formed by all flights, for most airlines their networks present the properties of the scale-free model. To find this result we considered the average degree and the standard deviation of the degrees, as well as the maximum degree, diameter, density and average path length. We find that these measures are consistent with scale-free graphs and the plot of average and standard deviation is inline with small world non random graphs, see figure 1. In the case of our data set, all the information of interest is in the connections, which we explore in the next section.

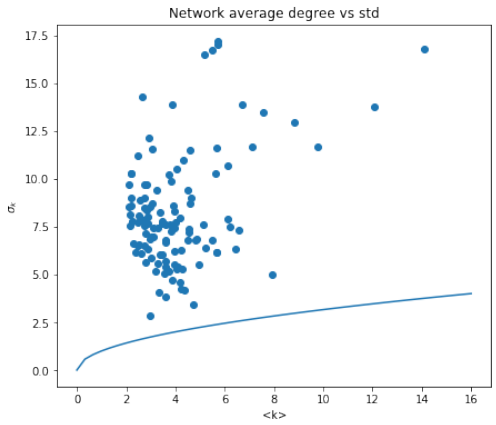

Fig. 1: average degree and standard deviation of graphs, bottom plot represents the relation a random graph would follow.

B. Airline Similarity Extraction

From the constructed graph, we extract a fixed number of features to numerically quantify the topology of the graphs. The features are:

- Algebraic connectivity of the graph: indication of connectivity and how the network we are analyzing is far from having isolated components. Whenever a network has isolated components, at least the second smallest eigenvalue is again 0. Results show that in a connected network, the gap between the second and the zero eigenvalues is a hint on the presence of strong clusters or not. [6].
- The multiplicity of eigenvalue 1 of the normalized Laplacian: indication of node duplications in a network. If large nodes in the graph might have all or many of their neighbours in common. [7]
- The spectral radius of the normalized Laplacian : empirical experiments have shown that the spectral radius of the normalized Laplacian is a good connectivity indicator of graphs with different size and order. We use this as an alternative to the algebraic connectivity.[7]
- The percentage of degree 1 nodes : the presence of many 1-degree nodes can heavily reflect the structure of a network when considering international flights, one airline can let you escape from its country but does not have internal flights in the foreign country.
- The percentage of nodes with degree larger than 5: airline network hubs play a big role in its robustness and connectivity.

To account for the different scales of features we standardized them to have zero-mean and unit-variance. As a first step, we consider only the algebraic connectivity of the graph and the multiplicity of eigenvalue 1. We re-scale the features using Principal Component Analysis. Principal Component analysis or PCA is a linear dimensionality reduction method that will return dimensions with the strongest variance. As we are working in 2 dimensions this just is a linear transformation of the space. The airlines plotted in this space are shown in figure 2.

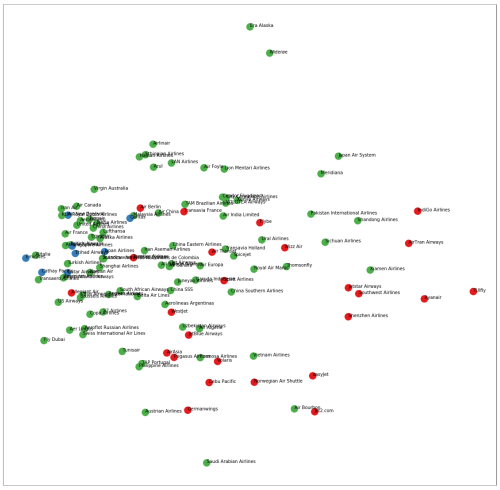

Fig. 2: clusters for airlines in topological space: red:low-cost airlines; green:regular airlines; blue:top airlines

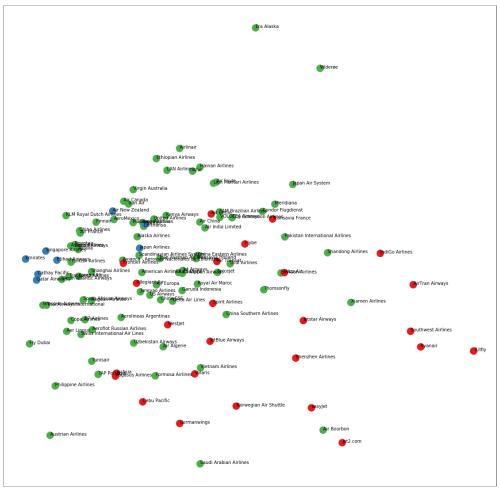

Fig. 3: clusters for airlines in topological space: red:low-cost airlines; green:regular airlines; blue:top airlines

As the algebraic connectivity is dependent on the number of nodes in the graph, we considered the node radius as an alternative measure. However, we found that for the task of assessing the similarity of graphs, the algebraic connectivity was better suited, as it showed higher variance among the points in space. As a next step, we additionally considered the percentage of degree 1 and degree larger than 5 nodes. As the features space is now larger than 3 we need to reduce the dimensionality. We do this by using PCA and reduce the features down to 2. Plotting these in these dimensions yields figure 3

C. Network Topology to Predict Delay

To give information on the effectiveness of network structure on the delay times of airlines we used a machine learning based approach. As we do not have average delay times for all airlines, we will used the subset of airlines for which we do have the associated data. We use a simple linear regression model. We measure the fit of the model with the R^2 measure, which indicates how correlated the predictions are with the original data. Using algebraic connectivity we find a weak correlation of 0.27, using the graph radius we find a slightly better 0.29 for the features tested. We will thus use graph radius going forward. Then, to increase the explanatory power of our model, as delays may be related to the robustness of the graph, we use the upper quantile and the median betweenness of nodes and edges. Node and edge betweenness are measures of the presence of bottlenecks in the networks. These measures themselves have a stronger explanatory power, with an R^2 score of 0.39. As a final experiment we combine the topological features and robustness measures. This increases our R^2 scores again to 0.61. This is quite good, given that we have assessed only the structure of the graph itself.

D. Finding Overlap in Airline Networks

We create an algorithm to compute an overlap score for two airlines. We take into consideration that two airlines with flights from the same airport to two different, but close-by airports, still have an overlap in their networks. Hence we not only use the airport source and destination, but also consider airports that are close to each other. In our algorithm, we have consider airports within 100 km distance between each other to be close. Our algorithm will iterate over all routes of the first airline, and check for matching routes for the second airline, increasing the score with 1 if such matches exist. Note that by using this method the overlap score is not symmetric. When designing the algorithm to compute the overlap scores, we encountered a Levenshtein distance-like algorithm for graphs, which roughly translates the number of changes one has to make to transform the first graph in the second graph. It measures how many changes do we need to make the graph isomorphic with each other. That does not work for our graph because the nodes are not interchangeable and are strictly tied to the geographical location.

E. Finding Potential for Collaboration

In order to establish which airlines may be interested in collaboration with other airlines we focus on how airline company B can help company A in providing new destination to A's clients. In short, we analyze weather B can enlarge the world-wide connection of A. For this task, we came up with a score that takes into account how many new destinations one company can reach when joining forces with another company. We call it $completionscore$ and it is equal to the number of new destinations airline B can provide to airline A when sharing its routes with A's clients. We can again notice that such score is not symmetric.

IV. Results

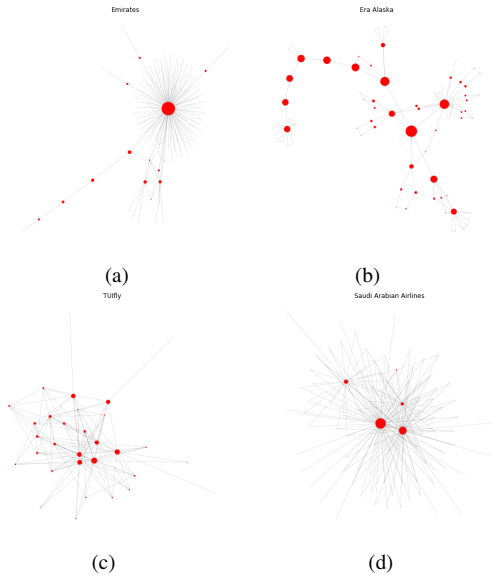

Fig. 4: Graphs with extreme value in two dimensional topology space. a) extremely positive x; b) extremely positive y; c) negative x; d) negative y

A. Results of the Topological Analysis

First, we analyze what the axes in the reduced feature space represent. Plotting the graphs fount at the extreme ends of the spectrum (which are the same in both features spaces), we see that the x axis corresponds to the number of hubs, while the y axis relates to how many nodes with similar position and degrees are in the graph we have 4. We find that the low-cost airlines are found within the same region and almost exclusively in that region. We can also see hints of top airlines being located all within the same region. We note that the extended features space performs better at this separation. In this representation the lower triangle (where the x axis $>$ y axis) contains 80% of all low-cost carriers, and 54% of the airlines in this region are low-cost carriers. Conversely, the upper triangle (where the x axis $<$ y axis) contains the remaining 20% of carriers, but these make up only 6% of all airlines in this region. This means that low-cost carriers have a strong tendency to share similar topology. The topology of these can be seen in figure 4c.

When considering top airlines, we again see that they are clustered together closely. Looking at their topology, we can see how they all have a very pronounced star-like

structure, and example of this can be seen in figure 4a.
The reason for this difference in structure lies in the operational model. Within the airline industry, these two topologies have been named point-to-point, referencing the distributed network structure of low-cost carriers and hub & spoke, referring to the hub-centric structure of traditional airlines [8]. Each of these models has their advantage. Hub Spoke allows for concentration of resources around a hub, and makes it easier to offer flights to many destinations by rerouting travelers through the hub. Point to point is popular with customers as many direct flights are offered. By only having direct flight this also allows to save money on infrastructure as luggage does not have to be rerouted and passengers do not have to transit. Additionally fuel costs are saved by direct flights.

Point-to-point is more suited for networks only offering short or mid-range flights, which is why most LCC do not offer long distance flights. We investigate the idea that the difference in network structures we observe is solely based on the difference between airlines that offer short, local flights and airlines that focus on intercontinental travel. However, checking for correlation between flight distance and how many international/intercontinental connections an airline serves, we only find small correlations of -0.09. Hence, the structure is not solely due to the destinations that an airline offers, as many local airlines also have the hub & spoke structure.
Keeping the two structures in mind we now consider the results from the delay analysis. We can observe a relatively strong relation between delay times and network structure, with betweenness and spectral radius of the normalized Laplacian playing the biggest roles as delay features. It is known that betweenness is one of the metric used to identify topology and fault-tolerance of networks. For the spectral radius, empirical results showed a relation with cTGD (compensated Total Graph Diversity): the measures tend to rank networks in a similar way [7]. The cTGD value is an indication of how well networks can survive when facing node or edge failures by redirecting traffic through other (convenient) existing paths [9]. We can indeed infer how such metrics have an impact on the delays of flights by highlighting how networks reflect more the Hub & Spoke or the star-like topology. The presence of hubs is a weakness for networks when coping to unpredictable events because they can accumulate traffic coming from many sources and fill out rapidly their capacities. Moreover, in a Hub & Spoke topology it is more difficult to redirect traffic because of the lack path diversity.

B. Finding Competitors and Potential Collaboration

From our analysis of topologies we also find that two northern airlines Era Alaska and Wideroe Airlines are

(a) (b)

Fig. 5: Airline graph of Era Alaska, resp. Wideroe Airlines

outliers. Analyzing their structure it is clear why. As can be seen in figure 5, they do not have the typical structure with a few hubs. They are most structured like a route network. This is due to these airlines replacing routes, as in very northern regions, air travel is used as an alternative to travel by car. Indeed, in both Norway and Alaska, air travel is an essential form of transportation [10], [11]. Considering countries, there is no obvious trend. For countries with most airlines, the UK, USA and China we find that they have airlines distributed all over the graph. The US and China have many close by airlines, which could indicate that there is in the market is quite crowded.

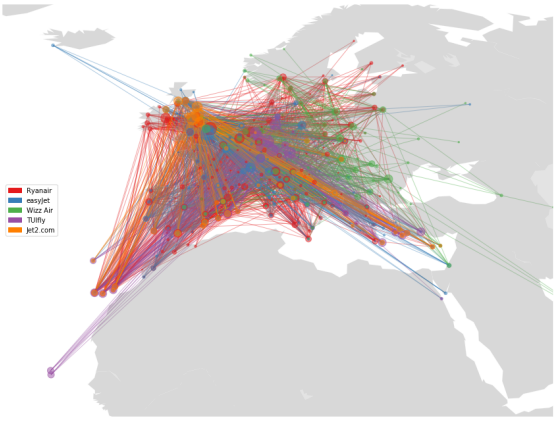

Fig. 6: European Budget Airlines with a strong point to point structure

Now that we know how similarity in topology is related to similar business models, we analyze airlines that share a strong topological similarity and see in how much competition they are to each other, as well as in how much competition low-cost and traditional airlines are to each other. This will give us an even deeper and more practical insight into the airline market.

For the purpose of this report we limit our analysis to a few airlines each, all with headquarters in Europe. Ryanair, Jet2.com and easyJet for the LCCs and Lufthansa, KLM Royal Dutch Airlines, Air France and British Airways for traditional (legacy) carriers. Plotting budget airlines and legacy airlines we can infer that their networks strongly overlap 6, 7. The algorithm we developed now helps us quantify this intuition. We find that while within a category

Fig. 7: European legacy airlines

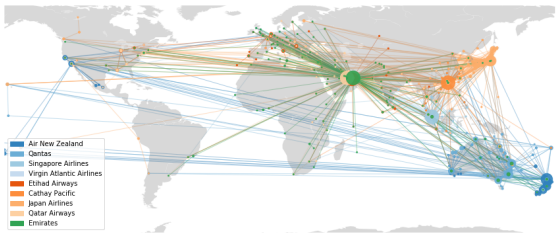

Fig. 8: Top-Airlines in the world with strong star-like structure

(LCC, legacy), there is quite some overlap indicating strong competition, airlines that are not in the same category to not show much overlap at all 9. This is not necessarily because these two kinds of airlines are not in competition, but because the routes they serve are very different.

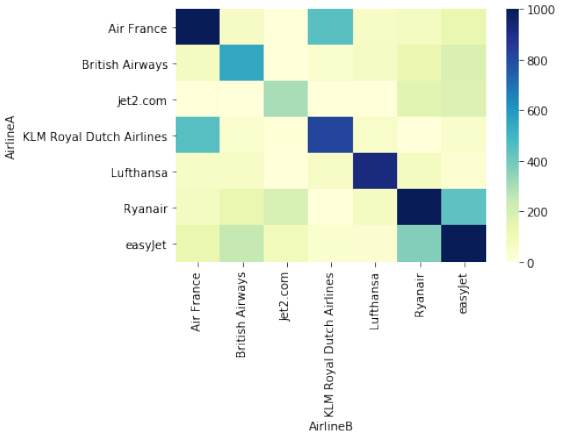

Fig. 9: Competition between large European airlines

To counter the competition, airlines may join up and form alliances to offer their passengers a wider option of places they can reach using that airline. The formation of these alliances may also be motivated on network topology. More and more traditional airlines have been changing their networks to move closer to a point to point style network structure, either to compete with LCCs or to form a more robust network. We check for partnerships within top airlines10. Top airlines all follow a similar business model cantering to business and more affluent travelers, making partnerships a more likely occurrence.

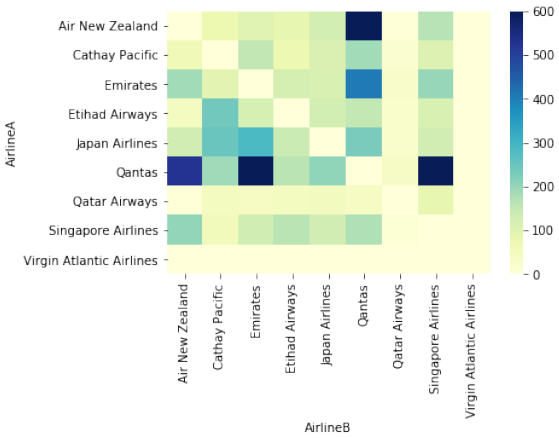

Fig. 10: Potential Benefit of Collaboration Between airlines

We can see that Emirates and Qantas, as well as Emirates and Japan airlines would get a huge benefit out of collaborating together. In fact, it turns out that there is collaboration between these two airlines and Emirates [12], [13]. This validates our method of collaboration finding.

V. Conclusion

Using graph theoretical measures we are able to not only gather insights into the business model of airlines but also understand why a common nuisance for passengers – delays – are more prevalent amongst some airlines. Extracting topological and robustness metrics we also show how they are correlated to the average delays one airline suffers when some unexpected event happens. Moreover, coupling such information with geographical locations of airports it is possible to reveal overlaps in airline routes and thus infer competitiveness. Additionally we are able to find interesting partnership opportunities and discover existing ones. All our measurements are based on theoretical analysis seen in class during the course and on other results found in the cited papers. Custom algorithms were also made on purpose for our analysis. Future works may consider analyzing other features related to an airlines operation using this topological approach.

References

[1] Jpatokal, "jpatokal/openflights," Jan 2019. [Online]. Available: https://github.com/jpatokal/openflights
[2] "List of low-cost airlines," Jan 2019. [Online]. Available: https://en.wikipedia.org/wiki/List_of_low-cost_airlines
[3] E. Rosen, "The 2018 list of the world's best airlines is out," Nov 2017. [Online]. Available: https://www.forbes.com/sites/ericrosen/2017/11/03/the-2018-list-of-the-worlds-best-airlines-is-out/63a0fcae5ed7
[4] "Airline list." [Online]. Available: https://airlinelist.co/
[5] "List of airports by iata code," Jan 2019. [Online]. Available: https://en.wikipedia.org/wiki/List_of_airports_by_IATA_code:_A

[6] U. von Luxburg, "A tutorial on spectral clustering," *CoRR*, vol. abs/0711.0189, 2007. [Online]. Available: http://arxiv.org/abs/0711.0189

[7] E. K. Cetinkaya, M. J. Alenazi, J. P. Rohrer, and J. P. Sterbenz, "Topology connectivity analysis of internet infrastructure using graph spectra," in *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2012 4th International Congress on*. IEEE, 2012, pp. 752–758.

[8] G. N. Cook and J. Goodwin, "Airline networks: A comparison of hub-and-spoke and point-to-point systems," *Journal of Aviation/Aerospace Education & Research*, vol. 17, no. 2, p. 1, 2008.

[9] J. P. Rohrer and J. P. Sterbenz, "Predicting topology survivability using path diversity," in *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2011 3rd International Congress on*. IEEE, 2011, pp. 1–7.

[10] "aviation in norway sustainability and social benefit," Jan 2008. [Online]. Available: https://avinor.no/globalassets/_konsern/miljo-lokal/miljoogsamfunn/sustainabilityandsocial-benefit.-summary.pdf

[11] "Rural alaska communities rely on this program for air travel. now it might be going away." [Online]. Available: https://www.adn.com/alaska-news/aviation/2017/04/09/rural-alaska-communities-rely-on-this-program-for-air-travel-now-it-might-be-going-away/

[12] "Japan airlines — our partners — emirates skywards." [Online]. Available: https://www.emirates.com/english/skywards/about/partners/airlines/japan-airlines.aspx

[13] "Qantas and emirates." [Online]. Available: https://www.qantas.com/travel/airlines/qantas-emirates/global/en