Prices in Air Transport*

Can Airport Network Characteristics Aid in Prediction?

Thor Donsby Noe[a], Morten Esketveit Rasmussen[a], and Christian Lund Sørensen[a]

[a]University of Copenhagen, Department of Economics, Denmark

May 24, 2019

Abstract

In this paper, we investigate whether network-related features of airports can add predictive power to a model predicting flight prices. We analyze the domestic US air transport sector as a network with airports as nodes, and direct flights as links. We find that the network is scale-free with a small subset of airports functioning as hubs in the network, and we produce a geographical map of the network with airports in their actual locations in the continental US for each of the years 1998, 2008, and 2018. Secondly, we find that while the network has changed substantially over the last two decades, the hub-and-spoke nature of the network has remained, and individual airport centrality has been relatively stable. Thirdly, we find that the network is vulnerable to removal of the airports acting as hubs, but fairly tolerant to removal of random nodes. Finally, we create a linear prediction model for flight prices to investigate whether network-related airport characteristics improve the predictive power in such a model. We employ elastic net regularization, and use K-fold cross-validation to determine the optimal hyper-parameters. We find that these features can be said to improve the prediction model only marginally, if at all.

*Access our Jupyter notebooks at `https://github.com/Morten-Esketveit/TSDS-gruppe-2019`

The author of each section is indicated by initial: Thor (T), Morten (M), Christian (C).

Contents

1. Introduction

Air transport facilitates commuting, trade, investment and tourism between and within countries. According to the US Federal Aviation Administration (Administration (2016)), aviation and the air transport sector in 2014 contributed 1,6 trillion dollars in total economic activity to the US economy and supported nearly 11 million jobs. Meanwhile, the network of airports connected through direct flights is a textbook example of a scale-free network characterized by a hub-and-spoke structure. In this paper, we firstly describe the basic characteristics of the US network of domestic flights for each of the years 1998, 2008, and 2018. We investigate the hub-and-spoke nature of the network including how the network is distributed geographically, assess the vulnerability of the network, and describe how the basic features of the network have changed over time. Secondly, we investigate whether network-related characteristics at the airport level contribute to predictive power in a model to predict flight prices.

Through characterization of the network we find support for a hub-and-spoke view of the network, with a minority of airports being exceedingly well-connected in the network while the majority of airports have direct flights to a relatively small number of airports. In other words, the network is scale-free, which is also apparent from the degree distribution of the network. We then take advantage of the geographical nature of the network, and produce a map of the network with nodes at their actual locations in the United States.

Furthermore, we conduct an analysis of how network characteristics are affected by the removal of certain nodes from the network, in order to assess the vulnerability of the network. In line with the results found in (Chi and Cai, 2004), we find that the network is vulnerable to removal of hubs (well connected airports), but reasonably tolerant to removal of random nodes and removal of the least connected nodes. Finally, we investigate how the network has changed over time.

The network analysis yields a number of airport-level characteristics that expand the feature set used in the second part of the paper where we attempt to predict flight prices using machine learning methods. The key question we wish to answer is, whether the network-related features add predictive power to the model? A priori, one might reasonably expect that the network characteristics matter for the airlines' planning of flights, and are informative of the competitive environment on a given route, both of which may affect flight prices. We find, that network characteristics only to an extremely limited extent improves the predictive power.

The remainder of the paper is structured as follows. In section 2 we briefly review related academic literature. Next, we provide the theoretical background in section 3, with an emphasis on the network theory that we draw upon. We then describe the data used in the analysis in section 4, before presenting the results of the analysis in section 5. In section 6 we discuss the results of the analysis, and we conclude the paper in section 7.

2. Literature review

2.1. US airline deregulation (T)

The classic air transport network through most of the 20th century consisted of simple point-to-point connections directly linking a small number of mayor cities as there were a relatively few number of flights overall (Martí et al., 2015). Delta Airlines had established their headquarter in Atlanta which became the busiest in the US as early as 1957[1], nonetheless, flying cross-country was still a complicated affair and it was not until the US Airline Deregulation Act of 1978 when cross-state competition was opened up and drastic changes came to the structure of the aviation industry (Forbes and Lederman, 2007; Daraban, 2012). Shortly after, the hub-and-spoke structure evolved both as a business model for the individual airlines and in cooperation with regional airlines (Forbes and Lederman, 2007) as a way of increasing the frequency and coverage. Furthermore, hub-and-spoke introduced economies of scale where administration and service was less important at the spokes as as most flights frequently pass through mayor hubs.

In contrast, Low Cost Carriers (LCC) emerged as well, offering point-to-point flights to secondary airports, often aiming to avoid transfers and expensive hubs (Daraban, 2012).

2.2. Fare determination (M)

The determinants of fares (flight prices) is investigated in a number of academic papers. In Vowles (2006), the author states that research has focused on three areas: The role of hub and spoke networks, airfare pricing determinants, and the role of LCC. In the aforementioned paper, the author investigates fare determinants in hub-to-hub markets. He finds, that the number of passengers, distance between airports, the share of low cost carriers, the presence of multiple airports in the geographical region and the presence of Southwest Airlines in the market or competing markets all have statistically significant effects on fares.

Transfers and stop-overs not only by definition add to distance travelled but also increase forgone time and fuel burn per mile travelled, ceteris paribus. In these two aspects hub-and-spoke networks are far less efficient than point-to-point flights due to transferring being both time- and fuel consuming with an extra set of landing, taxiing, waiting, and take-off as compared to cruising at altitude. Thus, while aviation fuel is exempted from taxes for now, simulating the development of aviation network given a carbon tax or a higher level of oil prices result in a higher share of direct point-to-point flights (O'Kelly, 2012).

Brueckner and Zhang (2001) presents a theoretical model of fare and frequency determination, and compares outcomes in a hub-and-spoke network to outcomes in a fully connected network. They find, that the hub-and-spoke network yields higher flight frequency, and higher prices for passengers whose origin or ultimate destination is a hub. The explanation for the somewhat surprising latter result is that higher frequency allows airlines to extract a higher fare from passengers in spite of lower costs.

[1] atlanta-airport.com/Airport/ATL/Airport$_{H}istory.aspx$

Abda et al. (2012) examines the impact of low cost carriers on fares in the United States. They find statistically significant effects of low cost carriers (entry or substantial growth) on flight prices.

2.3. Modelling Airport Networks (C)

Costa et al. (2011, pp. 41-42) provides a brief overview on the literature of networks in the context of airports including important findings and methodology. In this literature airports are defined as nodes, and flights as links. Typically these networks are treated as directed although most flights fly back and forth between two cities. Among the highlighted findings is that the degree distribution of airport networks seems to follow the power law (this is at least the case for studies of the World as a whole, India, Brazil, and the US), meaning that the networks is characterized by a few large hubs with a much higher degree than the average airport. It is further suggested that the topology of the network is primarily determined by GDP and population size across cities in the network.[2]

Chi and Cai (2004) show that the US airport network is affected by "errors" and "attacks". Their data set covers 215 US airports and flights during a specific week. To simulate an attack, they successively remove airports in order of importance, starting with the most connected airport. Conversely, to simulate an error, the authors remove airports in order of fewest connections. They then investigate how errors and attacks respectively influence key topological measures, specifically the average degree, the clustering coefficient, the diameter and efficiency. They find that these measures are affected far more by removal of the most well connected airports, compared to removal of the least connected airports.

Rocha (2017) outlines the fundamental properties of airport networks and surveys the literature on the dynamic modelling of these networks, that is, taking into account the time dimension. He divides the research on dynamic modelling into two categories; one in which long-term structural changes are analyzed using network “snapshots” and another that focuses on short-term changes which is used to analyze how delays propagate through the network and logistics of the airport industry.

Considering the first category, the literature mainly focuses on changes in basic network statistics over time. First, it is described how the deregulation of the US transportation sector changed the airport network from point-to-point to hub-and-spoke systems. Even though the airport networks have changed continually since then, the degree distribution seems to be quite constant. On the other hand, betweenness centrality and the clustering structure of the network seems to change over time.

Looking at the short-term dynamics, it is found that airports that may be close in the static network can be quite far away from each other in a dynamic network since some routes are operated very infrequently.

[2] Contrary to this, He et al. (2004) finds that the Chinese network of airports does not follow the power law.

3. Theory

In this section we briefly relate the market for commercial air transport to network theory. We first present basic characteristics of air transport. We present the network theory most pertinent to our analysis including the measures we use to both characterize the network and expand the feature set for the predictive model. The last subsection describes how we build a prediction model.

3.1. Airline business models

As is apparent in section 2, the common view within academia is that large airlines specialize in either of two mayor business models with very distinct network characteristics. That is, legacy airlines lean towards managing a hub-and-spoke network while low cost carriers (LCC) lean towards offering point-to-point connections. (Daraban, 2012; Baker, 2013; Martí et al., 2015).

3.1.1. Point-to-point connections (T)

Point-to-point connections are direct connections between two nodes. In aviation terminology it is routes connecting airports by direct flights as opposed to requiring transfers or layovers. The extreme being the fully-connected network in which every node is directly connected to each of the other nodes. While one can get anywhere in the network without wasting time and fuel by transferring in a hub, this also means that the required number of links, and thus carriers, grows exponentially with the number of nodes N such that the number of links L are (Bryan and O'Kelly, 1999):

$$L = \frac{N \cdot (N-1)}{2} \tag{3.1}$$

As noted by Barabási (2016) this will also be the maximum number of links in any non-directed network. That is, a complete graph of just 7 nodes would have no less than 21 links as seen in the left panel of figure 1 below.

3.1.2. Hub-and-spoke network (M)

The other extreme would be the single-hub or two-hub network where N nodes would only need $N-1$ connections as the N_s non-hub nodes, so called spokes, would only need to be connected to a hub in order to be indirectly connected to every other node in the network. O'kelly (1987) presented a hub-and-spoke network as a simple single-assignment model such that non-hub nodes only have one link, namely the one connecting them to a hub.

Given that all hubs are interlinked, drawing a few examples by hand will quickly reveal that the number of links L_{HS} in a stylized hub-and-spoke network depends on the number of hubs N_H and spokes N_S in the following way

$$L_{HS} = \frac{N_H \cdot (N_H - 1)}{2} + N_S, \qquad N_H > 0$$

As seen in the middle panel of figure 1 the two-hub network of 7 nodes only requires 6 links, however, the average shortest path length is 2 as opposed to 1 in the fully connected network (see paragraph 3.2.4.3).

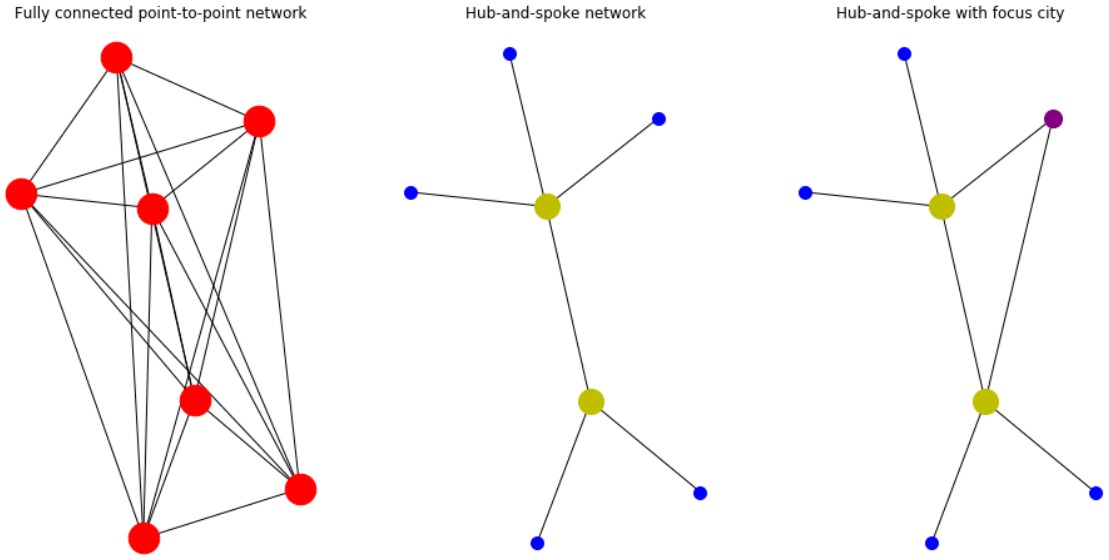
3.1.3. Focus cities (C)

A third business model can be added, namely the concept that smaller regional airlines mainly serve the so-called spokes (Forbes and Lederman, 2007). They usually serve as an integrated part of a hub-and-spoke network connecting the spokes to close-by hubs with smaller regional jets. However, regional airlines might also connect spokes to other spokes or even offering (infrequent) longer-range connections to other hubs, turning the spoke into a so called *focus city* or tiny hub (Mammarella, 2014).

Likewise some airlines are combining the two structures such as the LCC *Frontier Airlines* which first-and-foremost runs a hub in Denver. However, in 2012 they started experimenting with a new business model by expanding their service through offering long distance point-to-point flights to four popular West coast hubs via full-size jets from the small airport Colorado Springs (COS), just 90 miles (145 km) south of Denver International Airport.[3] This motivates us to choose COS as the airport for which we look into network changes over time in figure 3, section 5.1.

In the right panel of figure 1 the spokes in the upper-right corner has been turned into a focus city, thus, cutting off one transfer in order to get to one of the airports in the lower half of the network. While this can be seen as deviance from the very strict definition of the stylized hub-and-spoke network (O'kelly, 1987) we instead take on a looser approach in order to apply 'hub-and-spoke' as the best overall characterisation whenever we observe a network similar to the one we see in the right panel of figure 1.

Figure 1: Three different networks connecting 7 nodes



[3] http://gazette.com/springs-is-frontiers-new-front-in-battle-for-colorado-travelers/article/137275

3.2. Network Theory

We draw on network theory to characterize how airports are connected to each other and analyze what role the individual airport plays in the network of airports.

3.2.1. Air transport as a scale-free network (T)

A scale-free network is a network whose degree distribution follows a power law, i.e. $p_k \sim k^{-\gamma}$. One characteristic of the scale-free networks is the existence of 'hubs' - nodes that have a far higher degree than the average degree. Air transport is - quite literally - a textbook example of a scale-free network as seen in chapter 4 of Barabási (2016). In section 5 we will see that this is reflected in the degree distribution where most nodes have a low degree but some nodes (hubs) have a very high degree.

Given this *spoke-and-hub* nature of air transport, we expect the network to have some airports (hubs) that are connected to all or most other airports in its geographical vicinity, and well connected to similar hubs in other regions. This pattern may be repeating to some extent, insofar as there may be regional as well as national hubs.

We further expect the network to be sparse, i.e. that the number of links is (far) lower than the number of possible links. This also reflects the hub-and-spoke nature of the network, in that small local airports are likely to have a low number of links, particularly relative to the possible number of links. However, we expect the network to be connected, in the sense that, given an appropriate time frame, there will always be a path from one airport any other (through some number of other airports).

3.2.2. A Directed or Undirected Network (M)

Although individual flights are clearly directed, going from one airport to the other, we will primarily view the network as undirected. The reason being, that connections between airports are typically undirected.

As we will see in the data analysis, flights from airport i to airport j, almost always implies flights from airport j to airport i. We believe that this slight simplification to the network of airports is well suited for our analysis, that focuses exclusively on the effect on prices. An analysis of e.g. how delays propagate through the network, would require viewing links as directed and including time in a more intricate manner.

3.2.3. Weighted Network and Temporal Considerations (C)

We view the network as unweighted, that is, all links are considered to be equal. While this somewhat simplifies the analysis, we could have chosen to weight links e.g. by letting the number of flights between two airports determine the 'weight' of their common link.

The considered time frame clearly matters for the characteristics of the network, since not all flights between airports take place every day or even every week. Considering a years worth of flights will produce far more links in the network than considering the flights that take place on a particular day. We choose to analyze the yearly network, to ensure

that we also capture flights (links) that are seasonal in nature, e.g. flights that only happen in the summertime, or close to specific holidays.

3.2.4. Selected Network Measures (T)

In the course of the analysis, we draw upon a variety of measures to characterize the network. In the selection of measures, we look to Chi and Cai (2004). The measures we have chosen to use are presented below. They are: Average degree, clustering coefficient, diameter, average shortest path length and betweenness centrality.

3.2.4.1. Average Degree

The degree k_i of node i is the number of nodes with which node i share a link. In our case the degree of an airport is the number of airports to which it has a flight connection. The average degree is then simply the average across nodes. Average degree is calculated as:

$$\text{Average degree} = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N} \tag{3.2}$$

3.2.4.2. Clustering Coefficient

The clustering coefficient for a given node i is given as:

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{3.3}$$

Where L_i denotes the number of links between node i's neighbors. The clustering coefficient of the network is then given by:

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{3.4}$$

The clustering coefficient for a single node is equal to the fraction of possible links between node i's neighbors (nodes it is connected to) that are found in the network. A node whose neighbors are all connected to each other will have a clustering coefficient of 1. Conversely, if none of the nodes neighbors are connected, it will have a clustering coefficient of 0. Note, that for nodes with degree 1, the clustering coefficient is necessarily 0.

3.2.4.3. Average Shortest Path Length

The average shortest path length is defined as:

$$a = \sum_{s,t \in V} \frac{d(s,t)}{N(N-1)} \tag{3.5}$$

Where V is the set of nodes in the network, $d(i,j)$ is the length of the shortest path between nodes i and j, and N is the number of nodes.

3.2.4.4. Diamater

The diameter is the maximum shortest path length between any two nodes in the network.

3.2.4.5. Betweenness Centrality

Finally, we also calculate the betweenness centrality. For node i, this is the fraction of shortest paths between all nodes in the network that pass through node i (Brandes, 2008).

3.3. Prediction Model (M)

Based on the available data we want to assess whether network characteristics can improve a prediction model for flight prices. The simplest way of predicting a continuous variable is to use a linear regression model. We write our prediction model as

$$y_i = w_0 + WX_i$$

where y_i is the price of a given flight, w_0 is the bias/intercept, X_i is a vector of features, and W is a vector of corresponding weights.

To find the weights that makes the most accurate predictions we need to define a cost function. The standard cost function estimating a linear model is the **Mean Squared Error (MSE)**:

$$\text{MSE} = \frac{1}{N}\sum_{i}^{N}(y_i - (w_0 + WX_i))^2 \tag{3.6}$$

To address the problem of overfitting, we regularize the model using *elastic net*, which combines L1 ("Lasso") and L2 ("Ridge") regularization. Each of these regularization methods work by adding a term to the cost function. In the case of Lasso, the cost function then becomes:

$$J(w) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \alpha \cdot \sum_{j=1}^{p}|w_j| \tag{3.7}$$

where p is the number of features in the model. The hyper-parameter α then controls the strength of the regularization. In the case of Ridge regularization, the cost function becomes:

$$J(w) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \alpha \cdot \sum_{j=1}^{p}w_j^2 \tag{3.8}$$

Ridge regularization shrinks the weights in the model while Lasso tends to drive some coefficients to zero. Elastic net combines the two regularization methods, and adds an additional hyperparameter, the L1-ratio, that determines the relative importance of Ridge- and Lasso regularization. The model is numerically optimized using gradient descent.

3.3.1. Model evaluation (C)

In order to evaluate how our model performs on new data, we start by splitting the data into a training set and test set, keeping the latter untouched till the final model evaluation. The model is estimated using k-fold cross validated grid search. When estimating the model we are only using the training data. This allows us to get an unbiased estimate of the performance when we evaluate the model on the test data.

Estimating the model consists of two parts: one is to find the optimal weights in our prediction model - the other is to choose the optimal hyperparameters. The k-fold cross-validation randomly splits the training data into k folds, and then uses k-1 folds to train the model to optimize it before evaluating performance on the last fold. This procedure is then repeated k times, so that we obtain k models and the resulting score is an average of these k models. This k-fold cross-validation is done for each set of hyperparameter in the Cartasian product of the hyperparameters and the set of hyperparameters that performs the best are chosen. Finally, the model is evaluated on the test data.

4. Data and scraping

The data used in the analysis is comprised of data from multiple sources. We utilize three 'types' of data:

- *Data on airports.* In the network setting, these correspond to nodes in the network.
- *Data on flights.* These data sets contain information on flights between airports. As will be described in detail below, we have access to a rich data on US flights. In the network setting this information corresponds to links in the network.
- *Prices.* Data on prices is generally harder to obtain. For this reason, we have chosen to scrape web-data on prices from Skyscanner.

4.1. Airport Data (T)

Data on airports and comes from OpenFlights.[4] The dataset is being updated continuously and contains airport IATA code, name, city, country, geographical location (crs: WTS84), as well as a unique OpenFlights identifier, ICAO code and altitude. The dataset contains information on a total of 7,698 airports worldwide, of which 1,518 are located in the United States. Throughout the analysis, we focus on the airports that are found at least once in our dataset on flights, which is a total of 358 airports in 2018.

[4] Available at: `https://openflights.org/data.html`

4.2. Flights Data (M)

Flights data comes from the Bureau of Transportation Statistics, United States Department of Transportation[5]. The dataset covers all scheduled fligths by carriers (airlines) that account for at least one percent of domestic scheduled passenger revenues. To assess how the network has changed over time we collect flight data for 1998, 2008, and 2018. Each of the datasets include information on several million commercial flights within the US. For each flight, we have a range of information, including the aircraft carrier, origin and destination (IATA coded), distance travelled, time in air date etc. Flight data can be merged with airport data using the IATA codes on airports. The result is a network linking all US airports with at least one direct flight during that year.

4.3. Price Data (C)

We scrape the price data from skyscanner.com using the `Selenium` package for Python. The scraper function launches a browser and opens the URL for a given flight at our chosen date which is Monday, May 27$^{\text{th}}$ 2019. If possible, the price of the cheapest direct flight is chosen instead. The prices of flights operated by Southwest Airlines are not shown at Skyscanner. If their flight is the only one operating, the price of the cheapest indirect flight is chosen. The cheapest indirect flight is also chosen if no carrier is operating a direct flight at the date. For routes where we could not find a price this date, the script is rerun for the next day, the 28$^{\text{th}}$. In order to reduce the number of searches without result and thereby reduce time, we only choose routes with an average of more than one flight per week.[6]

There is a number of limitations to our approach. First, in order to reduce the scraping time, we are scraping prices for only 1 day. This implies that routes not operated on this particular Monday (nor Tuesday) are missing in the data set. Furthermore, choosing one specific date is problematic since prices show systematic changes across weekdays, and seasonality across the year. This approach is likely not to give an accurate measure of prices in general which may cause our prediction model to perform badly out of sample.

The price data set consist of prices for 5,283 of the total 6,425 routes. There is a number of routes that turn out to have extremely high prices because these routes are being operated by some sort of *air taxi* company. Modelling these prices is beyond the scope of this paper and therefore we remove them from the data set to end up with 5,083 observations.

5. Empirical Results

5.1. Network Characteristics (T)

When analyzing the network we utilize the `NetworkX` package for Python. We start by characterizing the network. Recall, that we view airports as nodes, and flights between

[5]Reporting Carrier On-time Performance 1987- (Accessed: 17-05-2019): `https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236`

[6]See Github for documentation and data: `https://github.com/Morten-Esketveit/TSDS-gruppe-2019`

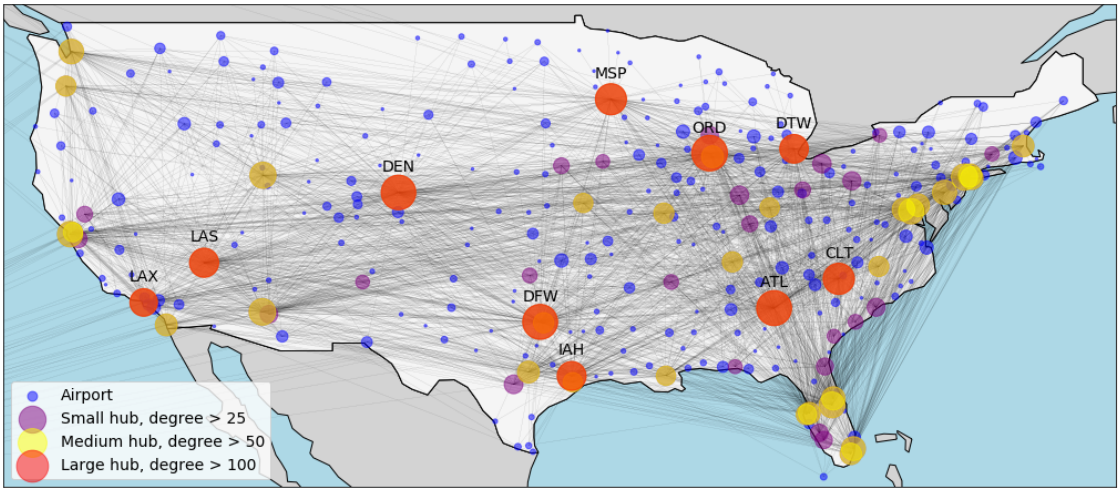
airports as links. Considering the data set of flights in 2018 from the Bureau of Transportation Statistics, we produce a network of US airports and the flights connecting them. We will refer to this network as the US Network. This network has 358 nodes and 3,110 links. Inserting the number of nodes N in equation (3.1), the maximum possible number of links would be:

$$L_{max} = \frac{358 \cdot (358 - 1)}{2} = 63,903 \tag{5.1}$$

This implies, that only approximately 5 pct. of possible links are actually found in the network. This sparseness of the network supports the hub-and-spoke view of air transport; rather than having all airports be connected, it is more economically feasible to have certain airports functioning as hubs in their geographical area, and connect to hubs in other regional areas.

Figure 2 below shows every airport in the network we consider, and the links that connect them. The network is distributed spatially according to the actual geographical locations of the airports, and overlaid on a map of the continental US. From a visual examination the hub-and-spoke nature of the network is clear. Hubs are somewhat distributed geographically and - at a glance - are placed near the larger population centres. Similar maps of the network in 1998 and 2008 can be found in figures A.1 and A.2 in appendix A. The network seems to have been relatively stable over time in the sense that airports that were well connected in 1998 are still well connected in 2018. The business model of a hub-and-spoke structure for the individual airlines is apparent as all of the 15 hubs for the two largest legacy airlines[7] are among the large and medium sized hubs, even more so, the network as a whole seem to resemble a hub-and-spoke structure.

Figure 2: Domestic flights network, 2018



To get a further sense of how the network is connected we calculate the *degree centrality*

[7] Ten hubs of American Airlines, `http://news.aa.com/multimedia/default.aspx#factsheets`
Eight hubs of Delta Airlines, `https://news.delta.com/corporate-stats-and-facts`

for each node in the network, i.e. the number of other airports each airport is connected to through flights in 2018. The average degree for the 358 airports contained in the data set is 18,0. This average masks a huge variation; the highest degree found in the data is the O'Hare International Airport (ORD) in Chicago with a degree of 176. Conversely, 51 airports have a degree of 1, implying that in this data set they only appear in connection with a single other airport.

The degree distribution for each of the years 1998, 2008, and 2018 can be seen in figure 5 and can also be grasped in figure 2 through the size and color of the nodes. Clearly, a large number of airports are connected by flights to few other airports, while a minority of airports are much more connected.

Table 1 shows key characteristics of the network for 1998, 2008, and 2018 respectively. For the network in 1998, average shortest path length and diameter is not defined, since the network is not connected. The number of airports has risen substantially in the period, as has the number of unique flight routes. The average degree of the network rose from 1998 to 2008, but fell slightly from 2008 to 2018. One should of course keep in mind, that the rise in the number of airports captures both construction of new airports and if existing airports are serviced by airlines that grow and then meet the inclusion criteria for the sample.

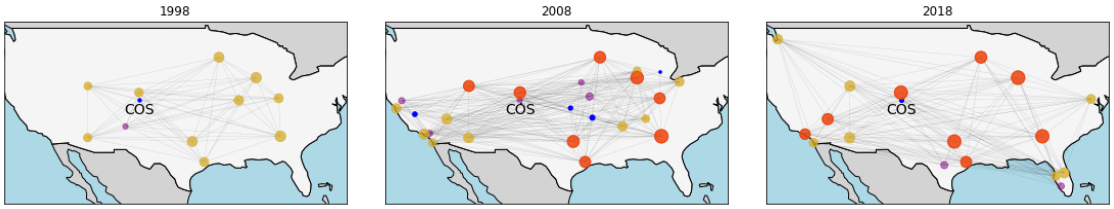
Table 1: Network Characteristics, 1998-2018

	1998	**2008**	**2018**
Nodes	209	310	355
Links	1614	2754	3110
Average degree	15.5	18.1	17.5
Average shortest path length	-	2.33	2.39
Diameter	-	5	6
Clustering Coefficient	0.63	0.65	0.57

To look into an example of how the network structure has changed over time we investigate the network consisting of the small airport Colorado Springs (COS) and its neighbors in figure 3. In the left panel we see that COS in 1998 was connected to 11 of the 10 closest inland hubs (except for Las Vegas) and only connected to a single airport that is not one of the absolute mayor hubs to an extent where the average degree between the 12 nodes is 10.8 out of possible 11, resulting in a *clustering coefficient* extremely close to unity. This is as close as one could dream of getting to observing a perfect hub-and-spoke network. In order to fly from COS to any West- or East Coast city (or vice versa) it is mandatory to transfer at one of the 10 bigger inland hubs. Interestingly enough, by 2008 several connections to smaller airports were maintained such that the average degree distribution only was 18.4 out of possible 26. This was followed by a reversion such that COS in 2018 again was almost exclusively connected to mayor hubs. The difference being that one now suddenly did not need to transfer to get directly to Florida or New York by the East Coast

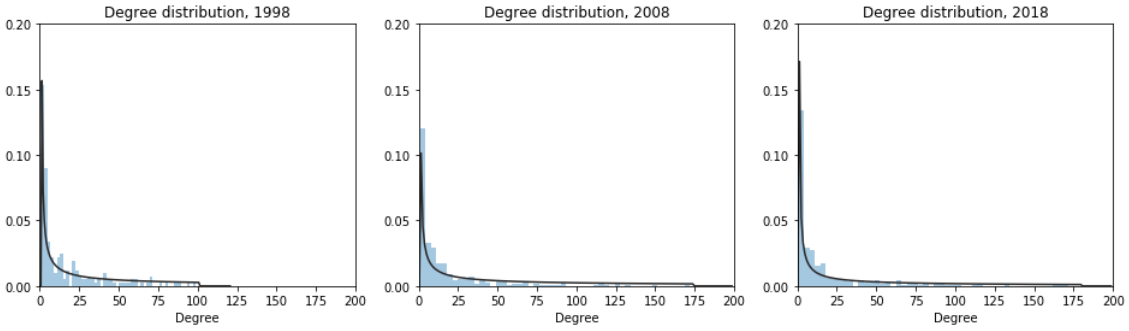
or to Las Vegas, LA or Seattle by the West Coast. Though the latter two possibly being down to *Frontier Airlines* as mentioned in subsection 3.1.3, the process of establishing Colorado Springs as a focus city with important long distance connections clearly did not stop there. However, we should refrain from extrapolating the development of the role of COS to being a general tendency in the network.

Figure 3: Network of Colorado Springs (COS) and its direct neighbors



The degree distribution of the network is shown in figure 5 with a fitted power-law distribution. The hub-and-spoke nature of the network - and the related fact that it is scale-free - is fairly apparent; a large number of nodes have a low degree, while a few nodes have a degree far above the average degree.

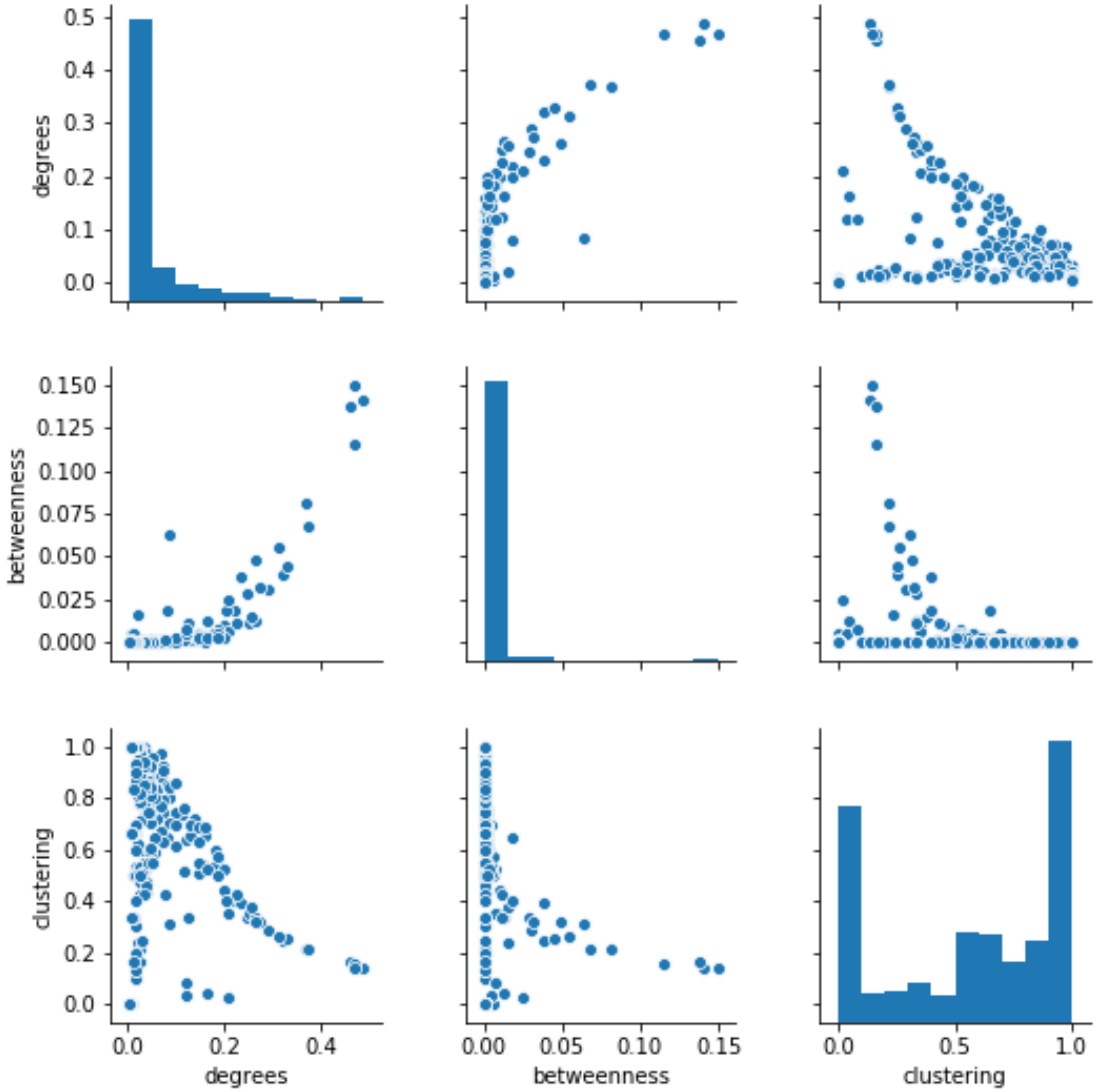
Figure 5: Degree distribution of network, 1998, 2008 and 2018



As an alternative measures of centrality in the network, we calculate *betweenness centrality* that measures the degree to which an node is part of the shortest path between two other nodes where path length is defined as the number of traversed links.

An overview of our three key network characteristics, and how they relate, can be seen in figure 6. First of all, there is a clear positive relationship between degree and betweenness centrality; airports that are linked to a large number of other airports are also more likely to be part of a shortest path between two airports. This is fairly unsurprising. Secondly, there is no simple relationship between clustering coefficient and either of the other measures. A large number of airports have clustering coefficient equal to 1, implying that all their neighbors are connected to each other. There is also a number of airports with a clustering coefficient of 0, implying that none of their neighbors are connected to each other, or that they only connect to one other airport.

Figure 6: Betweenness centrality, clustering coefficient and degree, 2018



In the end, we use clustering coefficient, betweenness and degree as features in the prediction model. Note, that these are all node-specific measures, and as such, for each flight observation, there are two realizations of each measure corresponding to destination and origin airport respectively.

5.2. Network vulnerability (M)

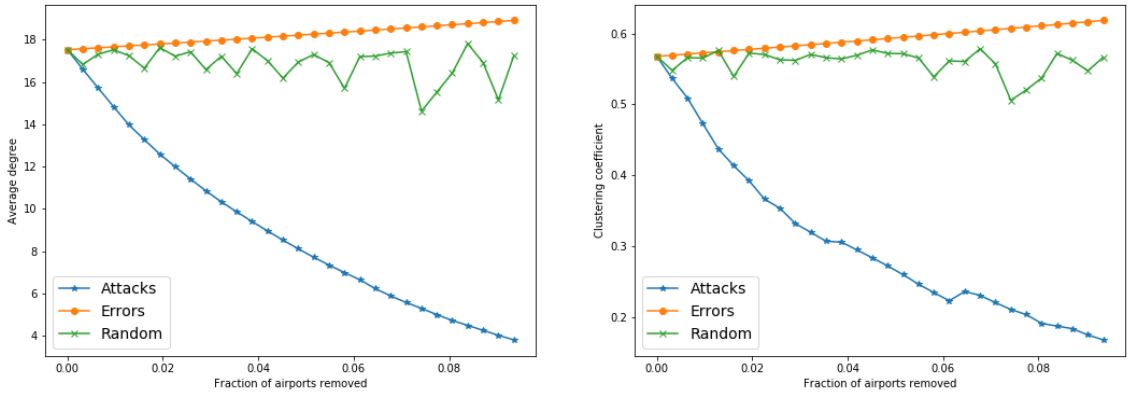
We conduct an analysis of how the network is affected when certain nodes are removed. This analysis is broadly in line with the analysis in Chi and Cai, 2004.

Figure 7 below shows how key network characteristics (average degree, clustering coefficient and global efficiency) are affected, when nodes are removed. One curve corresponds to removal of the most connected nodes, this is what we refer to as an 'attack' on the

network, where the hubs are targeted. We likewise consider how the network is affected when we remove the least important nodes first, the occurrence of a so-called 'error'.

It should be noted, that in this section we analyze initial consequences for a static network as it is in our data set. Obviously, if a number of airports were to close for a prolonged period (e.g. due to natural disaster), the agents in the market would make changes to the network to accommodate the new situation. Market forces are likely to have been a determinant in producing the network as it is, and exogenous changes to the network would cause changes in behaviour of the agents in the market which would produce a new network.

Figure 7: Effect of node removal on average degree (left) and clustering coefficient (right)



As figure 7 shows, the removal of a relatively small number of hubs can dramatically alter network characteristics. Unsurprisingly, removal of the least important nodes increases the measures while. Thus, the network is somewhat vulnerable to 'attacks' on the most connected airports.

In the figure, we also show the effect of removing random nodes. Note, that each point represents a "new" network of randomly chosen nodes in the original network. Nodes left out in the first observation may therefore be present in the second observation. Thus, removal of random nodes has close to no persistent effect on average degree or the clustering coefficient of the network due to mean reversion. This again points to the fact that the network is vulnerable only to a concerted attack on several important hubs.

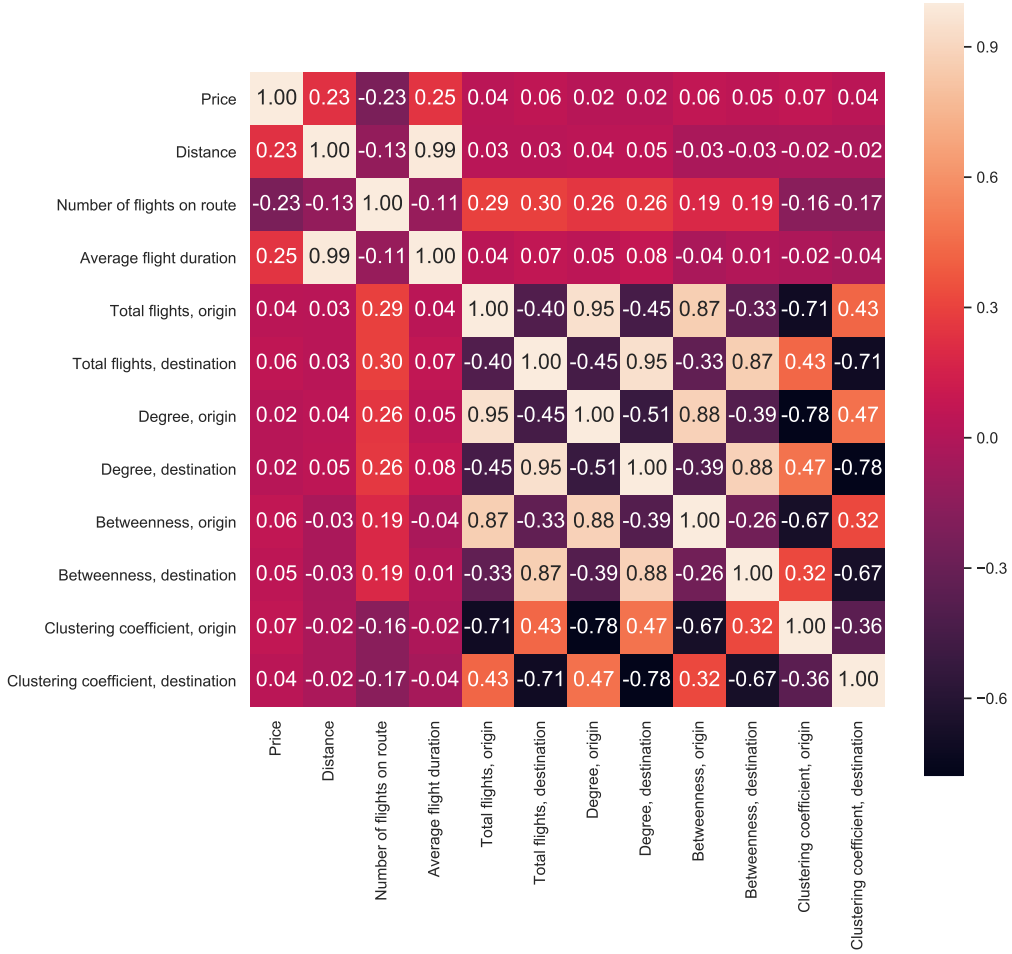
5.3. Predicting flight prices - Do network-related features add predictive power? (C)

We want to assess whether network characteristics can contribute in predicting flight prices compared to a prediction model using only a set of baseline variables. Our set of baseline variables include: The number of flights on the given route in the year considered, the distance between the two airports, the (average) time in minutes for the flight and the number of carriers who flew the route during the year considered, as well as dummies for origin and destination airports. Our set of network-related features includes, for both

origin and destination airport: Betweenness centrality, degree and clustering coefficient.

In order to get a sense of the correlation structures, we show all the continuous variables in the correlation plot in figure 8. Unsurprisingly, we see that prices are positively correlated with distance and flight duration and negatively with the number of flights on the route due to competition or economies of scale. We see that the rest of the variables, including all our network variables, are only weakly correlated with prices. It is further seen that some of the features exhibit multicollinearity to some degree.

Figure 8: Correlation Plot



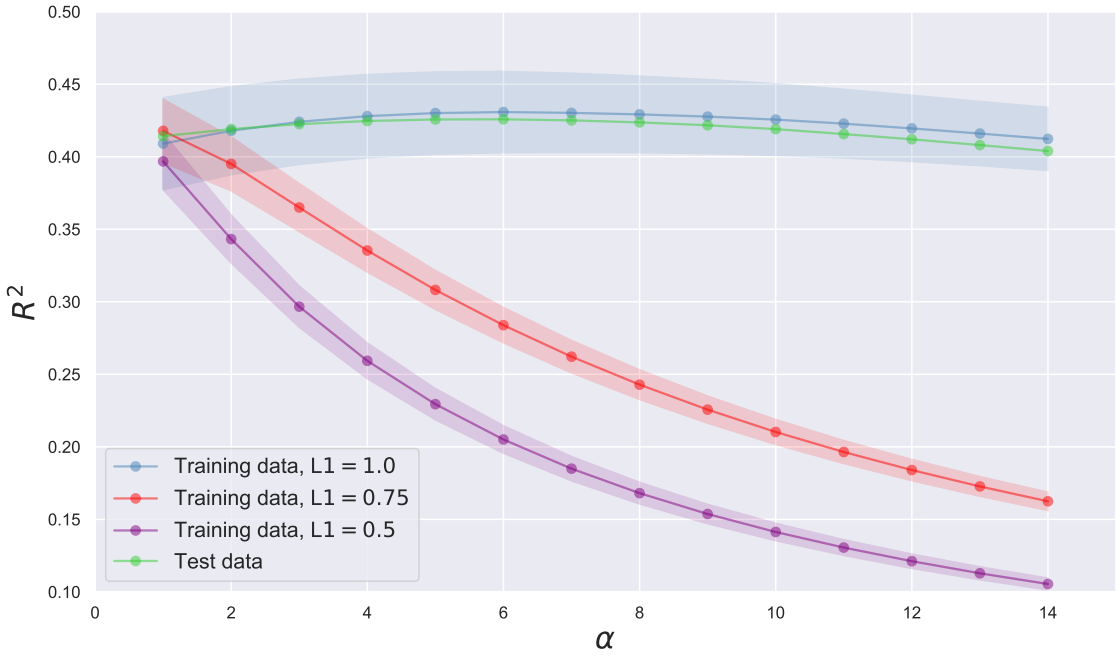
In order to predict prices we use a linear prediction model including network characteristics and other things such as distance, flight duration, and airport dummies as predictors. We follow the procedure outline in section 3.3. The procedure is implemented with the `SKLEARN` python module from which we use `train_test_split`, `StandardScaler`, `ElasticNet`, and `CVGridSearch`.[8]

We use a 5-fold cross-validation and for the grid search we allow the L1-ratio to vary

[8]See github for documentation, `https://github.com/Morten-Esketveit/TSDS-gruppe-2019`

between 0.25 and 1, and the alpha parameter to vary between 1 and 20.[9] Figure 9 plots the score, measured in terms of R^2 as a function of the regularization parameter α for the three highest L1 ratios. Using grid search the optimal α is approximately 6.4 whereas the optimal L1 ratio is 1.

Figure 9: Validation Curves



Finally, the best model is tested on our test data and the results are compared to the corresponding model without network features. The results are summarized in table 2. Our model is able to explain around 40 pct. of the variation in flight prices. We find, that adding network characteristics only increase the out-of-sample prediction marginally, i.e. by 0.5 percentage point. Considering the hyperparametrization both models have a L1 ratio of 1 whereas the α parameter differs slightly. Across the models, the weights of the baseline features are quite similar (see table 3 in appendix B) for the number of flights, average time, and the number of flights to and from the destination airport. These are the only continuous baseline features with weights that are different from zero. Among the other network features, origin betweenness and both clustering coefficients enters the model with non-zero weights. Based on the size of the coefficients, the clustering coefficient seems to be the most important, however the regularization tends to bias the coefficients. The way the network features enter the model is a result of the correlation structure showed in figure 8 and the Lasso regularization. Running the model without airport fixed effects (not shown) increases the importance of network characteristics - overall the prediction

[9] Prior to fitting the model, we scale the features using the `StandardScaler` in Python. This is done in order to avoid artificially large/small weights which could make convergence difficult when we are using regularization.

model performs much worse, however. In this case, the baseline model is able to explain 15.9 percent of the variation whereas the network model can explain 17.9 percent.

Table 2: Results, Prediction Model

		Baseline		**Networks**	
		Test data	Training data	Test data	Training data
Model evaluation	R^2	0.421	0.427	0.426	0.431
			(0.026)		(0.028)
Hyper parameters	α		6.359		5.872
	L1 ratio		1		1

6. Discussion (M)

Our network analysis is done with data from Bureau of Transport Statistics. While this data set is fairly comprehensive, it does not cover flights by airlines who account for less than one percent of revenues in the sector. The network is therefore not entirely 'complete' insofar as there may be routes or airports only used by airlines that do not appear in the data set.

The results of our analysis of the domestic US air transport network are broadly consistent with previous results about this network. Firstly, it is a scale-free network, where the majority of airports are linked to relatively few other airports while a few airports are very well connected and act as hubs. A very basic analysis and visual inspection suggests, that the network has become larger over time. Airports tend to keep their 'role' over time, although some airports became less well-connected in the period 2008-2018 while others became more well-connected. This may reflect changes 'on the ground'; cities and areas that experience e.g. economic growth may attract more flights from a larger number of airports. However, the process of how the network is 'generated' and changes over time is beyond the scope of this analysis.

Furthermore, the network is vulnerable to removal of hubs, but tolerant to removal of random nodes. This analysis of the vulnerability of the network is in line with previous academic work. However, it seems to be fairly hard to relate changes in network characteristics (e.g. average degree) for the air transport sector to how customers will actually be affected by a removal of these airports. The economic value of such an analysis is therefore questionable.

We have taken a fairly basic approach to predicting prices of flights between airports and this is reflected in the fact that we are only able to explain around 40 pct. of the price variation in our test data. There are a number of characteristics of the air transport sector that we do not include in our model, but that may conceivably raise predictive power. Firstly, due to large fixed cost the airport transport sector is characterized by imperfect competition. Airlines are price setters and prices therefore differ from the marginal costs. Whereas the marginal costs are closely linked to flight duration and distance, we have only

very limited proxies for demand such as the total number of flight on a given route within a year. In addition, price elasticities for demand are likely to vary across routes and time.

In relation to this, we have considered prices from only one Monday. Airlines tend to exert price discrimination depending on the day of the week. Some flights may have a reduced price on weekdays whereas others may have an increased price depending on whether the flight is considered a 'recreational' flight or a 'business' flight.

Moreover, airports are viewed as important infrastructure, and may represent large employers in their geographical area. Therefore, political concerns may cause interventions in the market that may affect prices on specific routes. Even more so, flight prices may also be affected by the types of planes servicing specific routes, and the airlines operating the routes. To incorporate such factors into the model would require a greatly enhanced data set, but would also likely increase the predictive power of our model.

As described in section 2, the presence of low cost carriers has been shown to affect flight prices. Including a variable to capture this is possible within our data set, but requires a systematic way of demarcating low cost carriers from other airlines. Finally, a large number of alternative node or link characteristics exist, some of which may contribute more predictive power in the model than the ones we have chosen.

With the above considerations in mind, the relatively poor performance of our prediction model is unsurprising. Our focus has been on investigating whether network characteristics contribute to predictive power. At a basic level, this also means that an attempt to build the best possible predictive model for flight prices has not been the primary focus. While this allows us to evaluate whether including network related characteristics contributes predictive power, it also implies that the conclusions we draw may not generalize. Network related characteristics may add predictive power in our model, but not in a more comprehensive model. This is clear from our analysis; network characteristics are much more important than when when applying a model without airport indicators to capture unobserved heterogeneity.

7. Conclusion

In this paper, we have have analyzed the air transport sector in the United States. The sector can be analyzed as a network of airports, connected by direct flights. In this network, a minority of airports are connected to a large number of other airports, and thereby function as hubs in the network. In other words, it is a scale-free network, following a hub-and-spoke structure. The network is vulnerable to removal of these key nodes, but tolerant to removal of either random nodes or the least well connected nodes. Furthermore, we find that the number of airports has risen steadily in the past two decades, but each airport broadly speaking performs the same function (hub or spoke) in the network over time. Specifically, airports that were well-connected in 1998 tends to be well-connected in 2018 as well. We produce geographical maps of the network in each of the three years considered, with nodes in their actual location in the continental US. The hub-and-spoke nature of

the network is clearly visible, and hubs seem to be fairly distributed geographically, but nonetheless placed near population centres.

To investigate the importance of network structure for prices we set up a linear model for predicting flight prices. Through grid-search cross-validation, we find, that the optimal hyper-parameters in our predictive model are an l1-ratio of 1, implying that a pure Lasso regularization is preferred to a linear combination of Lasso and Ridge regularization, and that the optimal α is around 6.4. Given our setup we find that network characteristics at the node level only very marginally, if at all, can be said to contribute to predictive power.

References

Abda, Mehdi Ben, Peter P Belobaba, and William S Swelbar (2012). "Impacts of LCC growth on domestic traffic and fares at largest US airports". In: *Journal of Air Transport Management* 18.1, pp. 21–25.

Administration, Federal Aviation (2016). *The Economic Impact of Civil Aviation on the U.S. Economy*.

Baker, David Mc A (2013). "Service quality and customer satisfaction in the airline industry: A comparison between legacy airlines and low-cost airlines". In: *American Journal of Tourism Research* 2.1, pp. 67–77.

Barabási, Albert-László (2016). "Network Science". In: Cambridge University Press. Chap. Graph Theory.

Brandes, Ulrik (2008). "On variants of shortest-path betweenness centrality and their generic computation". In: *Social Networks* 30.2, pp. 136–145.

Brueckner, Jan K and Yimin Zhang (2001). "A model of scheduling in airline networks: how a hub-and-spoke system affects flight frequency, fares and welfare". In: *Journal of Transport Economics and Policy (JTEP)* 35.2, pp. 195–222.

Bryan, Deborah L and Morton E O'Kelly (1999). "Hub-and-spoke networks in air transportation: an analytical review". In: *Journal of regional science* 39.2, pp. 275–295.

Chi, LP and X Cai (2004). "Structural changes caused by error and attack tolerance in US airport network". In: *International Journal of Modern Physics B* 18.17n19, pp. 2394–2400.

Costa, Luciano da Fontoura et al. (2011). "Analyzing and modeling real-world phenomena with complex networks: a survey of applications". In: *Advances in Physics* 60.3, pp. 329–412.

Daraban, Bogdan (2012). "The low cost carrier revolution continues: evidence from the US airline industry". In: *Journal of Business & Economics Research (Online)* 10.1, p. 37.

Forbes, Silke J and Mara Lederman (2007). "The role of regional airlines in the US airline industry". In: *Advances in Airline Economics* 2, pp. 193–208.

He, Yue, Xiangyang Zhu, and Da-Ren He (2004). "Statistics and developing model of Chinese skyway network". In: *International Journal of Modern Physics B* 18.17n19, pp. 2595–2598.

Mammarella, James (2014). "Encyclopedia of transportation: Social science and policy". In: SAGE Publications. Chap. Airport Hubs.

Martí, Luisa, Rosa Puertas, and Consuelo Calafat (2015). "Efficiency of airlines: Hub and Spoke versus Point-to-Point". In: *Journal of economic studies* 42.1, pp. 157–166.

O'kelly, Morton E (1987). "A quadratic integer program for the location of interacting hub facilities". In: *European journal of operational research* 32.3, pp. 393–404.

O'Kelly, Morton E (2012). "Fuel burn and environmental implications of airline hub networks". In: *Transportation Research Part D: Transport and Environment* 17.7, pp. 555–567.

Rocha, Luis EC (2017). "Dynamics of air transport networks: A review from a complex systems perspective". In: *Chinese Journal of Aeronautics* 30.2, pp. 469–478.

Vowles, Timothy M (2006). "Airfare pricing determinants in hub-to-hub markets". In: *Journal of Transport Geography* 14.1, pp. 15–22.

A. Appendix: Domestic flights network, 1998 and 2007

Figure A.1: Domestic flights network, 1998

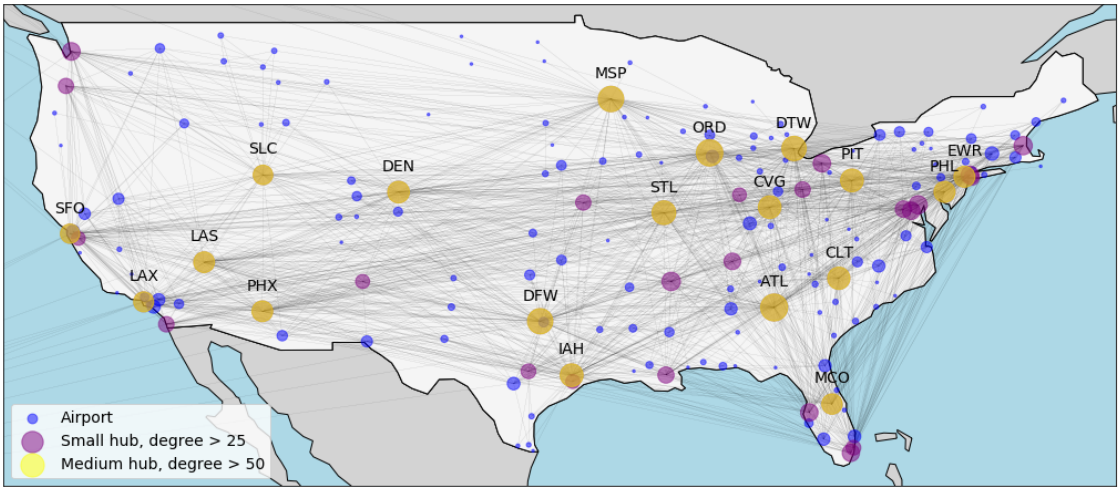
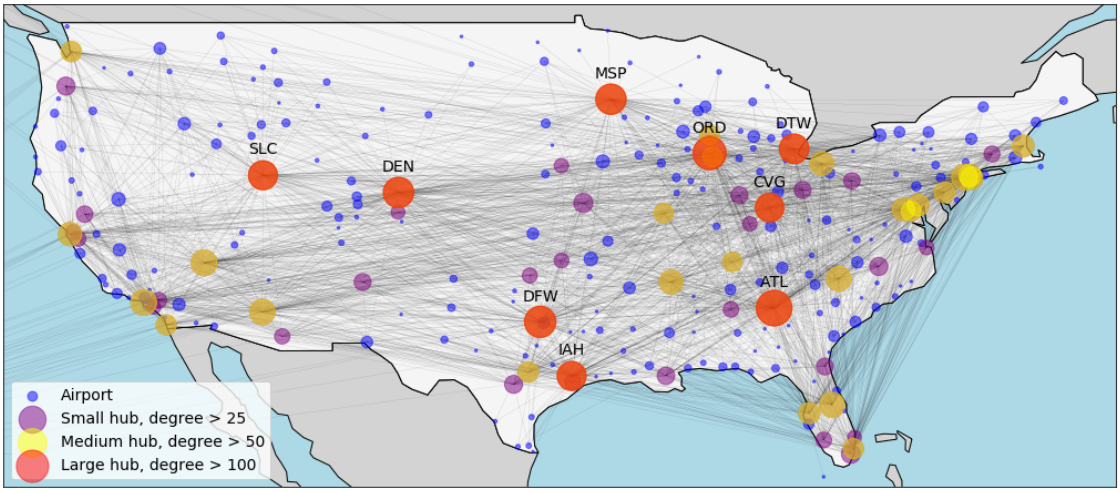


Figure A.2: Domestic flights network, 2008



B. Appendix: Prediction model

Table 3: Model coefficients

Variable	Baseline model	Network model
Distance	0.00	0.00
count	-160.62	-149.12
avg_time_mins	233.13	244.41
Origin_flights	0.00	0.00
Destination_flights	16.40	21.87
origin_degree		0.00
dest_degree		0.00
origin_btwns		26.84
dest_btwns		0.00
origin_clustcoef		81.56
dest_clustcoef		30.99
Airport FE	✓	✓
Company FE	✓	✓