Dynamic Networks: with Application to U.S. Domestic Airlines*

David de Jong, ENAC and Delft University of Technology
Jesse Hagenaars, ENAC and Delft University of Technology
Aalok Parkash, ENAC and Delft University of Technology

(supervised by Steve Lawford, ENAC)

January 18, 2016

Abstract

We investigate the network structure and dynamic behaviour of airline carriers operating on U.S. domestic routes over the years 1999 to 2013. We use graph theory to describe global and local features of the networks and show (a) that these measures can be used as explanatory variables in regression models, and (b) that dynamic analysis of the measures can provide insight into subtle changes in network structure over time. Furthermore, we build regression and machine learning models, and use empirical analysis, to study fare-setting and route-entry.

*We are grateful to Nathalie Lenoir for supporting this project. Correspondence can be addressed to Steve Lawford at LEEA, ENAC, 7 avenue Edouard Belin, CS 54005, 31055, Toulouse, Cedex 4, France; email: steve_lawford@yahoo.co.uk [Lawford] or daviddejong@me.com [de Jong] or jessehagenaars@gmail.com [Hagenaars] or a_parkash@msn.com [Parkash]. The usual caveat applies. JEL classification: C2 (Single Equation Models, Single Variables), C88 (Other Computer Software), D85 (Network Formation and Analysis: Theory), L14 (Transactional Relationships, Contracts and Reputation, Networks), L93 (Air Transportation). Keywords: Airlines, centrality measures, machine learning, networks, Python.

1 Introduction

In today's world, many different networks can be observed. Consider your social network, the railroad network or (a bit more complicated) the network of neurones inside your brain. All these networks have different characteristics that tell us about their size, how well connected they are, and how they evolve over time.

By studying a network, one can learn a lot about the interactions between the nodes or participants of the network. For example, Robinson and Stuart [1] argue that the structure and size of alliance agreements between firms in the biotechnology sector is affected by past alliances formed by these firms. They show that when two firms have a large network of past alliances, their deal is less likely to include the purchase of shares, indicating greater mutual trust.

Looking at the personal network of CEOs, El-Khatib et al. [2] show that CEOs that have a more prominent position within their network of business professionals have better access to private information and have greater bargaining power concerning mergers and acquisitions. Also, they are more likely to pursue acquisitions, and these deals are more likely to destroy equity value.

Hochberg et al. [3] study the investment performance of venture capital (VC) firms. They found that VC firms with better networks have significantly better investment performance.

Considering airline networks, a lot of similar interactions and characteristics can be observed. For example, the structure of the network: is it point-to-point or hub-spoke? How large is the network? How do carriers compete on routes? Does a monopoly position translate into higher fares? A particularly interesting phenomenon in the world of airline networks is the so-called 'Southwest Effect'. According to Goolsbee and Syverson [4], the Southwest effect can be described as the significant cutting of fares by incumbent airlines following Southwest's entry (or threat of entry) on routes in the U.S.

In order to define the aim of this paper clearly, the research questions are stated below.

- How can airline networks be defined?
- What phenomena can be observed from the interaction between fares and other characteristics of a route?
- Can the 'Southwest Effect' be confirmed?
- How does competition affect fares?
- Can the evolution of airline networks be predicted?
- Can mathematical network simulation be used to simulate 'real world' airline networks?
- Can Machine Learning be used for classification predictions on network evolution?

First of all, Chapter 2 will explain the source of the data and its composition. Then, the first research question will be dealt with in Chapter 3. The second research question will be answered in Chapter 4. Chapters 5, 6 and 7 will cover the third, fourth and fifth research question, respectively. Finally, Chapter 9 will conclude.

2 Data

2.1 Source

The data used in this paper is based on the DB1B and T-100 databases from the U.S. Department of Transportation, covering domestic routes from 1999Q1 to 2013Q4. The reason for choosing domestic U.S. flights lies in the fact that these databases are available free of charge and are very extensive.

For this paper, the dataset was limited to non-stop round-trip coach-class tickets, which were then aggregated to non-directional tickets on the route-carrier-quarter level. To signify this data reduction: the original raw databases had a size of 150 GB, whereas the one used in this paper is 'only' 189 MB. It consists of 102,526 route-carrier-quarters, which are defined by keys of the following format: ABE_ATL_DL_1999_1, for the route from Lehigh Valley International to Hartsfield-Jackson Atlanta International, operated by Delta Airlines in 1999Q1. To conclude, the data set covers a total of 37 carriers serving 231 airports.

2.2 Carriers of Interest

The dataset used for this paper contains 37 carriers. However, for most of the research, these will be limited in order to simplify the analysis.

2.2.1 AA (American Airlines)

American Airlines is the world's largest airline, both by fleet size and revenue. It is a major airline which operates on both a domestic and international network. AA is also a founding member of the Oneworld airline alliance. As of June 2015, AA has a fleet of 953 aircraft and mainly operates Boeing and Airbus aircraft. On top of this, American has over 350 aircraft on order from Airbus and Boeing, which is the largest fleet renewal in its history.

Since its formation in 1930, American Airlines has merged with numerous carriers throughout the years. It is interesting to note that American itself was also formed by a merger of 80 carriers. The most recent merger, in 2015, was with US Airways. This merger would form the largest airline group in the world. After the merger, the combined airline would continue to carry the American Airlines name and branding, but the US Airways management team would retain most operational management positions. Also, US Airways would have to give up their spot in the Star Alliance, whereas American Airways would stay in the Oneworld Alliance.

2.2.2 AS (Alaska Airlines)

AS is a major airline and part of the Alaska Air Group, along with its sister airline Horizon Air. The airline is known for having attained the highest customer satisfaction level among the traditional airlines for eight consecutive years.

Alaska is not in any of the major airline alliances. Instead, it has codeshare agreements with some of the members from Oneworld and Skyteam. The airline was founded in 1932 as McGee Airways where it flew a single engined, three-passenger aircraft. As of today, Alaska has a fleet of 147 aircraft, all of which are Boeing 737 aircraft.

2.2.3 B6 (JetBlue)

JetBlue Airways Corporation is an American low-cost carrier and the fifth largest airline in the United States. The carrier was founded in August 1999 under the name 'NewAir'. It started in the same style as Southwest, by offering low-cost tickets, but also tried to distinguish itself from Southwest by offering services such as in-flight entertainment, TV at every seat and satellite radio. Currently, JetBlue also offers a business class section in some of

its aircraft used for transcontinental flights. JetBlue is not a member of an airline alliance. As of October 2015, its fleet consists of Airbus and Embraer aircraft. The operating base of the carrier can be found at JFK.

2.2.4 DL (Delta Airlines)

DL is a major airline which has its headquarters at the world's largest hub: Hartsfield-Jackson in Atlanta, Georgia. Delta is one of the founders of the Skyteam alliance and it is the oldest airline in the United States. Delta operates a fleet of more than 800 aircraft, with the largest Boeing 767 and Airbus A330 fleet of any US carrier. An interesting difference between many US airlines and Delta is that Delta utilizes many older aircraft. They believe that this is the way to have the highest profitability.

2.2.5 F9 (Frontier Airlines)

Frontier is a low cost carrier in the US and operates flights to 50 domestic destinations. The airline has a hub at Denver International Airport. Frontier is quite a young airline, as it was incorporated in 1994. Frontier has a fleet of 61 aircraft, all of which are manufactured by Airbus. The airline also has a codesharing agreement with Great Lakes Airlines. By doing this, the airline is able to connect passengers through Denver and Phoenix to the surrounding Rocky Mountain States.

2.2.6 FL (AirTran Airways)

FL is a low-cost carrier located in Dallas, Texas. Previously known as ValuJet airlines this company started operations in 1993. In 2014 they merged with Southwest. The main advantage Southwest had when they bought AirTran in 2010 is that AirTran had slots at Atlanta airport. By merging with AirTran, Southwest was able to gain entry to the largest airport in the world as ranked by number of annual passengers.

2.2.7 NK (Spirit Airlines)

Spirit Airlines is a low cost carrier founded in 1980 as Charter One. It is based in Florida. On March 6, 2007, Spirit began a transition to an ultra low-cost carrier. Their initial plan was to begin charging US$10 per checked bag for the first two bags, $5 if bags are reserved before 24 hours prior to the flight, in addition to charging $1 for drinks which were previously complimentary. Spirit Airlines is currently the only airline in the US with a 2-star rating from Skytrax. They currently serve 57 destinations and own a fleet of 78 aircraft. In 2014 they generated almost 2 billion dollars in revenue.

2.2.8 SY (Sun Country Airlines)

Sun Country is an airline carrier based in Minneapolis. They operate scheduled flights and charter flights in the US, Mexico and the Caribbean. The company commenced operations in 1982 and is still active today. Although it went bankrupt in 2001 it returned to activity. The carrier lost a big part of its fleet when it went bankrupt, and therefore decided to keep a uniform fleet consisting of solely Boeing 737-800s. Currently, Sun Country owns 20 planes and serves 35 destinations.

2.2.9 UA (United Airlines)

United is the world's largest airline when measured by the number of destinations. It is also a founding member of Star Alliance, the world's first and largest global airline alliance. The airline is a legacy (or major) carrier with a large domestic and international network. Regional service is carried out under the name United Express. As of

November 2015, the (mainline) fleet is large, with 720 aircraft, and consists mainly of Boeing aircraft, ranging from 737 to 787 types, with some Airbus A319 and A320 planes. United's three largest hubs are Chicago O'Hare, George Bush Intercontinental Airport (Houston) and Newark.

United was founded in 1926 as an air mail service. Over the years it has merged/acquired several other airlines, most recently in 2010, when it merged with Continental Airlines. After the 2001 terrorist attacks, in which two United aircraft were involved, and the following airline industry downturn, United Airlines went bankrupt in 2002. After cost-cutting and recovery operations, the airline emerged from bankruptcy in 2006.

2.2.10 US (US Airways)

US Airways was a major American airline that ceased operations in 2015, when it merged with American Airlines. Until 2014, the airline was part of the Star Alliance, after which it joined Oneworld in preparation for the merger with AA. As of April 2015, its fleet comprised 331 aircraft, made up mainly of Airbus aircraft, ranging from A319 to A330, with some Boeing 757s and Embraer 190s. The airline's two largest hubs were Charlotte Douglas International Airport and Philadelphia International Airport.

US Airways was founded in 1937 as an air mail carrier under the name All American Aviation. In 1949 it switched to passenger transport and was renamed All American Airways.

Before merging into American Airlines, US Airways had plans to be acquired by United Airlines in 2000. However, after objections from different parties, United withdrew from the deal. In the follwing years, US Airways twice filed for bankruptcy, despite severe cost-cutting. US Airways merged with America West Airlines in 2005 and made a bid for Delta Airlines in 2006, but that was opposed.

2.2.11 WN (Southwest Airlines)

Southwest is the world's largest low-cost carrier. It was established in 1967 and, as of 2014, it carries the most domestic passengers of any US airline. Southwest has not been part of any alliance. Its fleet consists only of Boeing 737 aircraft, making it the world's largest operator of this aircraft type. Being a low-cost carrier, Southwest does not make use of the hub-spoke network concept, but uses a point-to-point network.

Over the last couple of years, Southwest has lost some of its low-cost attributes. It engaged in codesharing, started serving primary airports and attracted business travellers through frequent-flyer programs and other extra services. In 2010, Southwest acquired AirTran Airways. Through this deal, the airline eliminated one of its competitors on the low-cost market, while expanding its network to Mexico and the Caribbean.

2.3 Variables

The database contains the following variables for each route-carrier-quarter:

- **origin:** Origin airport of the route.
- **destination:** Destination airport of the route.
- **carrier:** Carrier operating the route.
- **year:** Year in which the route is operated.
- **quarter:** Quarter in which the route is operated.
- **T100airtime:** In-flight time of the route, in minutes.
- **T100loadfactor:** Average load factor of the route, in percentages.
- **T100pax:** Total number of passengers transported.
- **T100ramptoramptime:** Ramp-to-ramp time of the route, in minutes.
- **T100seats:** Total number of seats available.
- **WNpresence:** Indicates whether Southwest is present on the route or not.

- **absTempDiff:** Absolute temperature difference between origin and destination, in Fahrenheit.
- **destinationLatitude, destinationLongitude, originLatitude, originLongitude:** Latitude and longitude of the origin and destination.
- **distance:** Distance covered by the route, in miles.
- **hhiDB1B:** The Herfindahl index, which gives an indication of the competition on a route. A value of 1 indicates a monopoly, whereas a lower value means that more than one airline has some market share on that route.
- **largeAirport:** Indicates whether one of the airports of the route is a large airport.
- **marketShareDest, marketShareOrigin:** Market share of the carrier at the destination/origin airport of the route.
- **marketShareRoute:** Market share of the carrier on the route.
- **maxGDP:** Max gross domestic product across route, so either the destination or the origin.
- **maxPopulation:** Max population across a route.
- **meanFare:** Average fare on a route.
- **meanGDP:** Average gross domestic product across a route.
- **meanGDPperCapita:** Average gross domestic product per capita.
- **meanPopulation:** Average population across a route.
- **meanRealFare:** Average fare, corrected for inflation to 2013Q4 prices.
- **p10Fare, p10RealFare, p25Fare, p25RealFare, p50Fare, p50RealFare, p75Fare, p75RealFare, p90Fare, p90RealFare:** Fare percentiles, both with and without correction for inflation.
- **pax:** Average number of passengers per flight.
- **time:** Quarter counter, with 24 being 1999Q1 and 83 being 2013Q4.

3 Network Measures

This section introduces mathematical tools for describing (airline) networks.

3.1 Graph Theory

Real-life networks can sometimes be very difficult to analyze mathematically. In order to do so effectively, one often models these networks as graphs. An example of a graph can be seen in Fig. 1.

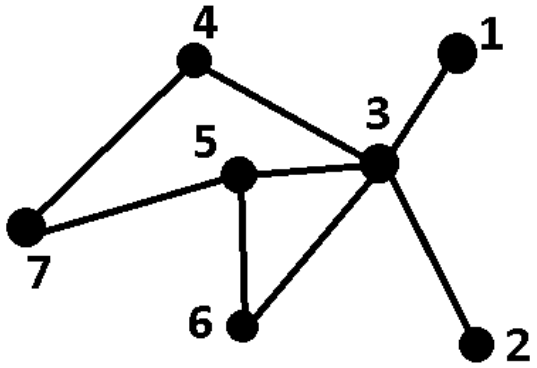

Figure 1: *An example of a simple graph.*

The study of graphs is called *graph theory*. Below, some of the key concepts of graph theory will be explained. This, as we will see later, is of great use since graph theory was used extensively throughout the research.

- **Nodes:** The numbered dots that can be seen in Fig. 1 are referred to as nodes or vertices. In real life, these can represent various objects that interact with each other. In this research of the domestic airline network of the US, the nodes represent the airports.
- **Edges:** The lines that connect the nodes are called edges or links. In this research, the links represent the routes that carriers fly between the airports. Note that the edges in Fig. 1 originating from a node do not directly link back to the same node. In other words, there are no self-links. In this research that is also true, because an airplane will never depart from an airport, and land at that same airport (without landing somewhere else first).
- **Directed/undirected:** A graph can be either directed or undirected, depending on the types of links in the graph. In the case of Fig. 1 it can be seen that every node can interact with another node as long as it is linked to it. Sometimes, however, this interaction can only take place in one direction. In that case, the link is replaced by an arrow pointing in the direction of the interaction. Because of this, we say that the graph in Fig. 1 is an undirected graph.
- **Degree:** The degree d_i is specified for a particular node i, and is the number of edges incident to it. For example, the degree of node 5 is equal to three, or: $d_5 = 3$. In the case of a directed graph a distinction is made between the indegree d_i^- and outdegree d_i^+. The indegree is the number of incoming links, and the outdegree refers to the number of outgoing links.
- **Walk/path**: A walk is defined as an alternating path of nodes and edges, whereas a path is the same but without repeating nodes.
- **Adjacency matrix:** Graphs are very illustrative, but in order to perform statistical analysis on these graphs, it is often important to model the graph mathematically. This is done in the adjacency matrix. The adjacency matrix is an n by n matrix, where n is the number of nodes present in the graph. It consists only of 1's and 0's, with a 1 if there is a connection between two nodes, and a 0 if there is no connection. The adjacency matrix of the graph in Fig. 1 is shown in Table 1.

Table 1: *The adjacency matrix corresponding to the graph of Fig. 1.*

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	0	1	0	0	0	0
2	0	0	1	0	0	0	0
3	1	1	0	1	1	1	0
4	0	0	1	0	0	0	1
5	0	0	1	0	0	1	1
6	0	0	1	0	1	0	0
7	0	0	0	1	1	0	0

It is interesting to note that the adjacency matrix in Fig. 1 is symmetric. This immediately shows that its corresponding graph must be undirected: if there is a link (1) between nodes 1 and 3, there will always be a link between nodes 3 and 1. The same logic applies for no link (0).

3.2 Macro Statistics

There are several measures that can be used to define graphs, divided into macro and micro statistics. Macro statistics will be covered in this subsection, while micro statistics will be dealt with in the next subsection.

- **Connectedness:** A network can be called connected if every pair of nodes has an 'edge' or connection between them. In total there are 235 airports in our data set. After the adjacency matrix (g) for a given route-carrier-quarter is built we remove all the isolated 'nodes' or airports from this matrix. This leaves us with an adjacency matrix which contains all the airports of the given route-carrier-quarter. If not every pair of nodes in a network is linked it is called *disconnected*.

- **Diameter:** The diameter of a network is defined as the maximum number of links that will be have to be crossed when 'travelling' between any two nodes in a network while taking the shortest route. Caution needs to be taken when analyzing this macro network parameter. The DB1B database only specifies the origin and destination. No layover data is provided. Therefore it can not be determined if a passenger travels from A to B with or without layovers at other airports.

- **Density and average path length:** The density of a network is defined as the actual number of links or 'edges' over the total number links. This statistic takes a value between 0 and 1. If this statistic has a high value it means the carrier is using a point-to-point network. A low value usually indicates a hub-and-spoke network. The average path length is defined as the average of all the shortest routes in between any two nodes pairs in a network. If the density and average path length in a network are both low it is safe to say the carrier utilizes a hub-and-spoke network strategy.

- **Degree distribution:** The degree distribution is $P(d)$ is the fraction of nodes with degree d. The degree of a node represents the number of connections that a node has to other nodes. For a hub and spoke network it can be expected that there are relatively few nodes with a high degree (the hubs) and relatively high amount of low degree nodes (the spokes). The degree distribution $P(d)$ can be calculated as follows:

$$P(d) = \frac{1}{n}\sum_{i=1}^{n} 1(d_i = d), \tag{1}$$

  Where $1(*)$ is an indicator function.

3.3 Micro Statistics

- **Degree centrality:** The degree centrality gives an indication of how connected a node is in terms of direct links. It is defined as:

$$DC_i(g) = \frac{d_i}{n-1}, \tag{2}$$

  where d_i is the number of airports that node i is connected to. It takes values between and 0 and 1, meaning it is either isolated or linked to every other node respectively.

- **Closeness centrality:** The closeness centrality is the inverse of the average shortest path length $l_i j$ between a node i and all other nodes j. It is defined as:

$$CC_i(g) = \frac{n-1}{\sum\limits_{i:i\neq j}^{n} l_{ij}} \tag{3}$$

The farther away node i is, the lower the closeness centrality becomes. A relatively high value of this statistic usually points to a hub-and-spoke network.

- **Betweenness centrality:** The betweenness centrality measures the importance of node i in connecting nodes j and k to the rest of the network averages across all pairs of nodes. This can be described mathematically as:

$$BetweenC_i(g) = \frac{\sum_{k<j:i\neq j,i\neq k}^{n} \frac{P_i(k,j)}{P(k,j)}}{\frac{1}{2}(n-1)(n-2)}, \tag{4}$$

where $P_i(k,j)$ are the number of routes that between node i and j that go via node i, and $P(k,j)$ is the total number of routes between node j and k.

- **Eigenvector centrality:** This is defined as:

$$\lambda\, EigenvectorC(g) = g\, EigenvectorC(g), \tag{5}$$

where λ is an eigenvalue and $EigenvectorC(g)$ is a right-hand eigenvector. Therefore, the importance (centrality) of a node is determined by the importance of the nodes that are directly linked to node i. By convention, the eigenvector corresponding to the largest eigenvalue is used.

To conclude, airline networks can be described in terms of a graph with nodes and edges, on which a number of macro (global) and micro (local) statistics can be defined.

3.4 Application of Network Measures

The network measures that have been introduced above can be applied in many different settings. In order to relate the network measures to the field of interest, the U.S domestic airline network, some of the most important network measures have been computed for the observed networks of Southwest Airlines (WN) and Frontier Airlines (F9).

Figure 2: *Networks of WN (left) and F9 (right)*

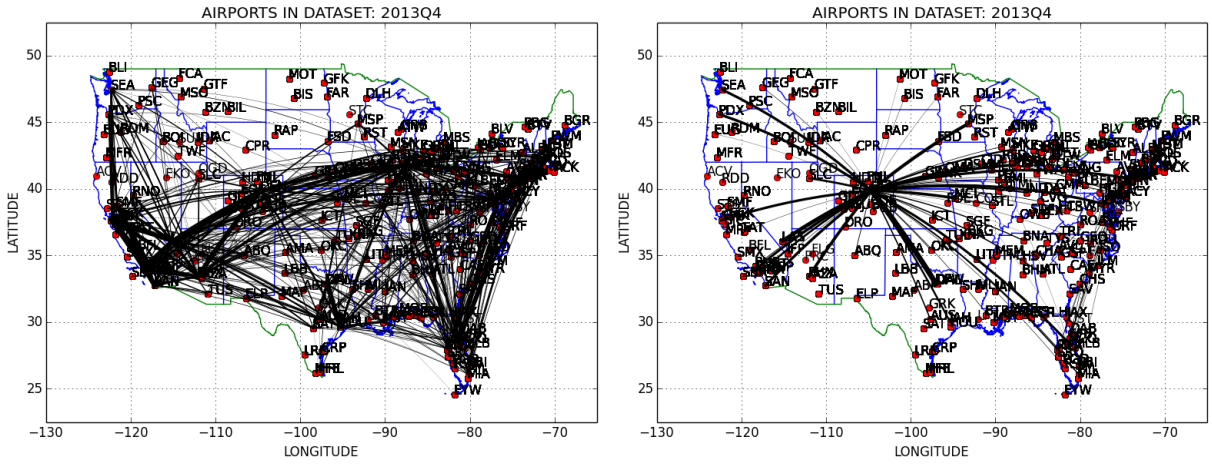

Table 2: *Measures of the networks shown in figure 2*

	# nodes	# edges	diameter	density	BC	CC	DC	EC
WN	88	522	3	0.14	0.12	0.73	0.62	0.27
F9	58	70	4	0.04	0.96	0.92	0.95	0.68

When looking at the networks shown in Fig. 2, it can immediately be seen that the network of Southwest is very dense and has a point-to-point network with some small concentrations at Denver, the south-west and the east coast. Frontier on the other hand is not quite as dense and is a textbook example of a hub and spoke network, with a big hub at Denver International Airport.

Keeping these characteristics in mind, it is very interesting to look at the statistic results shown in Table 2. From the nodes and edges we can see that Southwest serves routes to slightly more airports (nodes) than Frontier , but the number of routes (edges) that it serves is much larger. The difference is density is also quite significant. In general, a density of 0.14 is considered to be quite high for airline networks.

The four centrality measures are measured at Denver International Airport, which is a big hub for Frontier, and a smaller base for Southwest. The most interesting centrality measure is the Betweenness Centrality (BC) which is much larger for Frontier than it is for Southwest. If we recall the definition of BC, it is the importance of a node in connecting two other nodes. This is obviously the case for Denver in the network of Frontier because it is such a large hub. If you want to go from one place to another, you will most likely have to go through Denver. Another interesting centrality measure is the Degree Centrality (DC) which describes the connectedness of an airport. This measure is also significantly higher for Frontier because it is a large hub.

4 Modelling of Networks

In this chapter, the following research question will be answered: 'What phenomena can be observed from the interaction between fare and other characteristics of a route?' In order to gain more knowledge about the data set we decided to use a linear model. Using the statistical programming language R, a model was built that consisted of a linear relation between *meanRealFare* and a particular set of control variables. By determining the coefficients of these variables, we learned a great deal about the data and about certain relationships between the variables.

4.1 Model 1

4.1.1 Model Characteristics and Variables

In order to improve the quality of the model, several variables were added to the data set. They are based on the passenger market share on a route and indicate whether, on a route, there is a monopoly, a duopoly or competition. In more detail:

- **monopoly:** A route is considered a monopoly if a single carrier captures more than 90 % of the passenger share.
- **duopoly:** A route is considered a duopoly if the sum of the passenger shares of the two largest carriers exceeds 90 %.
- **competitive:** All routes that are neither a monopoly, nor a duopoly.

Furthermore, centrality measures were included in the data set. These were determined for origin and destination, as well as the maximum and minimum of each centrality measure on a route.

First, a linear model was built with the control variables *distance*, *meanGDPperCapita*, $T100loadfactor$, *absTempDiff*, *marketShareRoute*, *monopoly*, *duopoly*, *mindegree* and *maxdegree*.

$$meanRealFare = \alpha + \vec{\beta} \cdot \vec{x}_{control} + u. \tag{6}$$

Here, α is the intercept and β the coefficients for the various variables, and u is the error. Table 3 summarizes the estimated model.

Table 3: *Summary of the linear model with control variables distance, meanGDPperCapita, T100loadfactor, absTempDiff, marketShareRoute, monopoly, duopoly, mindegree and maxdegree.*
*Significance codes: *** 99.9 %, ** 99 %, * 95 %, . 90 %, blank 0 %.*

	Coefficient	Significance
intercept	386.0	***
distance	0.1507	***
meanGDPperCapita	-0.9118	***
T100loadfactor	-3.384	***
absTempDiff	-0.4221	***
marketShareRoute	117.4	***
monopoly	0.8119	
duopoly	7.238	***
mindegree	-164.2	***
maxdegree	116.4	***
adjusted R^2	0.356	N/A

4.1.2 Explanation of the Coefficients

The coefficient for the *distance* variable is positive, which means that the fare increases as the distance increases, a perfectly sensible result. The significance code indicates that the effect of the distance on the mean real fare is very significant.

The variable *meanGDPperCapita* has a negative value, which indicates that fares are higher for 'poorer' routes. What could explain the negative sign is the fact that 'richer' routes have better infrastructure, which reduces cost and which could lead to cheaper tickets. Fares for 'poorer' routes will have to compensate for higher costs due to worse infrastructure.

The control variable $T100loadfactor$ also has a negative sign, indicating that fares will be higher for routes with a lower load factor. This makes sense, since carriers will want to maximize revenue when the load factor for a route is not high, therefore charging higher fares. On routes with higher load factors, fares can be lower.

The second control variable, *absTempDiff*, has a negative sign as well. A higher temperature difference between origin and destination will thus be coupled to a lower fare. This could be explained by the fact that a high temperature difference is likely to be a leisure route instead of a business route. Fares on leisure routes are often lower than on business routes due to the demands of business travelers.

The last control variable is *marketShareRoute*, which has a positive sign. This indicates that, if a carrier has a larger market share, the fares it charges will increase. This makes sense, since a large market share gives a carrier a more dominant position on a route, resulting in power to increase fares. This for example happens with a monopoly.

For the *monopoly* variable, the coefficient is positive. This would indicate that on a monopoly route, the fares are higher compared to a competitive route. This result makes sense. However, the significance code indicates that the influence of the *monopoly* variable on the *meanRealFare* variable is very insignificant and can be ignored. Still, it is included in the model since it acts as a control variable.

Same as with the *monopoly* variable, the positive coefficient for the *duopoly* variable means that on a duopoly route, the fares are higher compared to a competitive route, which also makes sense. This time however, the significance code indicates that the effect of a duopoly on the mean real fare is significant.

The *mindegree* variable has a negative coefficient, indicating that the minimum degree on a route centrality decreases as the fare increases. In other words, if the least connected airport on a route become less connected, the fare increases. It works the same the other way around. This can be explained in two ways. First, the revenue of an airport decreases as the number of connections decreases, therefore higher fares have to be charged to make up for that loss. Second, as the least connected airport will become less connected, it will start to look like the spoke of a hub-spoke network. Due to the hub effect, fares from and to hubs are higher, which most likely explains the increase in fares in this case.

The *maxdegree* variable has a positive coefficient. Therefore, fare increases when the most connected airport on a route becomes more connected. This makes sense, again explained by the hub effect, that states that fares on a hub are higher due to the dominant position of a carrier.

4.2 Model 2

4.2.1 Model Characteristics and Variables

A second model was built in order to further examine the hub effect, especially over time. This was done by multiplying the *mindegree* and *maxdegree* variables by the $time$ variable. To summarize, the linear model is:

$$meanRealFare = \alpha + \vec{\beta} \cdot \vec{x}_{control} + u, \tag{7}$$

where α is again the intercept and β the coefficients for the various variables $x_control$, which are: $distance$, $meanGDPperCapita$, $T100loadfactor$, $absTempDiff$, $marketShareRoute$, $monopoly$, $duopoly$, $mindegree$, $maxdegree$, $mindegree * time$ and $maxdegree * time$, and u is the error. The calculation of the coefficients is given in Table 4.

Table 4: *Summary of the linear model with control variables distance, meanGDPperCapita, T100loadfactor, absTempDiff, marketShareRoute, monopoly, duopoly, mindegree, maxdegree, mindegree∗time and maxdegree∗time. Significance codes: *** 99.9 %, ** 99 %, * 95 %, . 90 %, blank 0 %.*

	Coefficients	Significance
intercept	307.8	***
distance	0.1428	***
meanGDPperCapita	0.4440	***
T100loadfactor	-2.440	***
absTempDiff	-0.3182	***
marketShareRoute	110.7	***
monopoly	6.771	**
duopoly	7.425	***
mindegree	-513.3	***
maxdegree	207.0	***
time	-0.9245	***
mindegree*time	6.696	***
maxdegree*time	-1.959	***
adjusted R^2	0.380	N/A

4.2.2 Explanation of the Coefficients

Most coefficients have the same signs as in the previous model. The ones that have changed will be explained, along with the new variables.

The variable *meanGDPperCapita* now has a positive value. This indicates that for 'richer' routes, fares are higher. This makes sense, since carriers can maximize their revenue by charging higher fares on these routes. On the other hand, on 'poorer' routes, demand can be kept high by charging lower fares.

For the $time$ variable, the sign is negative, indicating that fares decrease as time passes: a very sensible result, since flying becomes cheaper and cheaper due to decreasing costs.

The $mindegree * time$ variable has a positive coefficient, which means that, over time, the influence on the fare of the increasing 'hubness' of one of the airports on a route due to the least connected airport becoming less connected becomes more significant. This could be explained due to the fact that there are a lot of airports that could become less connected since, for that, no additional infrastructure is needed.

The $maxdegree * time$ variable and its negative coefficient can be explained in a similar way to the $mindegree * time$ variable. The negative sign indicates that, over time, the effect on the fare of the 'hubness' of one of the airports on a route due to the best connected airport becoming more connected becomes less significant. For a large airport to become even larger, a lot of additional infrastructure is needed, for which there is often no space. Also, it is very hard to increase capacity by filling more slots, since most hubs are already very congested.

4.3 Combining Results

To conclude this section, some of the explanations will be highlighted in order to answer the research question. It was defined as follows: 'What phenomena can be observed from the interaction between fare and other characteristics of a route?'

- Fares are higher on routes between 'richer' cities
- Fares are higher on routes with lower load factors
- Fares on business routes are higher than on leisure routes
- Fares on monopoly and duopoly routes are higher than on competitive routes
- Fares increase as a route becomes more and more the spoke of a hub-spoke network

5 The 'Southwest Effect'

The research question that will be answered in this section is as follows: 'How can the 'Southwest Effect' be confirmed?' According to Goolsbee and Syverson [4], the 'Southwest effect' can be described as the significant cutting of fares by incumbent airlines following Southwest's entry (or threat of entry) on routes in the U.S. In order to confirm this effect, data on fares will have to be examined.

This was done in two ways. First, a single route where Southwest made entry after some time was compared with a route where they did not enter. Second, for each carrier the difference in fares for route-quarters with Southwest presence was compared to route-quarters without Southwest presence.

For the first part, the route from Denver International (DEN) to Phoenix Sky Harbor International (PHX) was compared to the route from DEN to Dallas/Fort Worth International (DFW). On DEN-PHX, Southwest entered in quarter 28 (2006Q1), while it never entered DEN-DFW. Fig. 3 shows the evolution of the fares over time and the entry of Southwest.

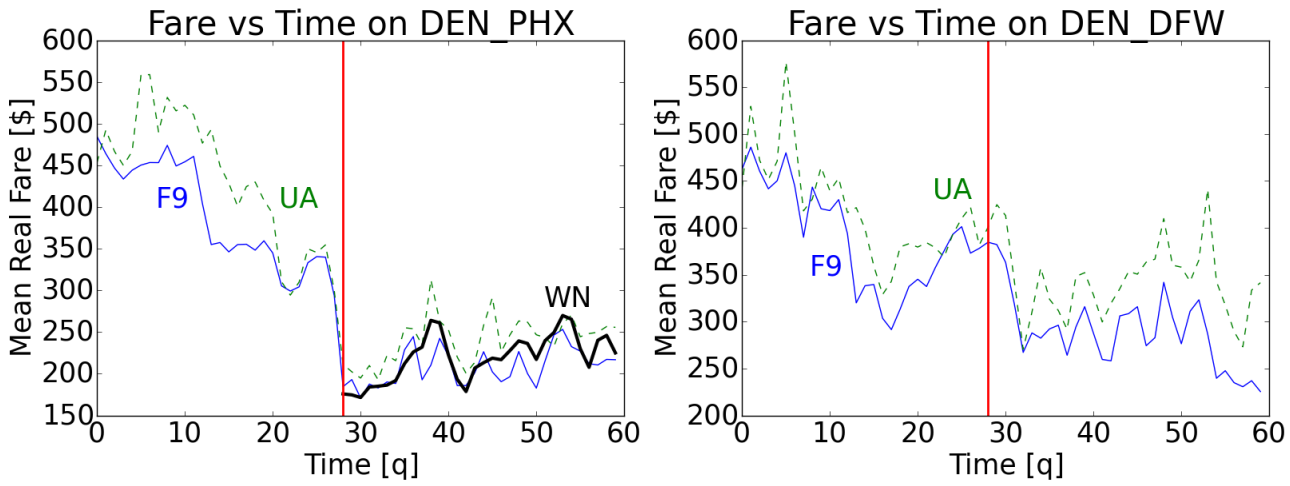

Figure 3: *Fares over time on routes DEN-PHX (with Southwest entry) and DEN-DFW (without Southwest entry).*

As can be seen clearly from Fig. 3, the incumbent airlines on route DEN-PHX, Frontier (F9) and United (UA), start to cut fares significantly in the period just before entry of Southwest, anticipating its arrival. When comparing between the quarter before entry and the quarter of entry (2005Q4 and 2006Q1), the decrease in fare for United is 29.5 %, while the decrease in fare for Frontier is as large as 37.2 %.

Looking again at Fig. 3 but now at route DEN-DFW, where Southwest does not enter, no decrease in fares takes place during the same two quarters (2005Q4 and 2006Q1). In fact, fares increase by 5.0 % for United and 1.7 % for Frontier. This suggests that the fare-cutting on DEN-PHX is not due to some general fare-cutting strategy, but really due to the Southwest entry.

For the second part, the difference in fares for route-quarters with Southwest presence was compared to route-quarters without Southwest presence. This was done for each carrier. The results are given in Tab. 5.

Table 5: *Fare difference between route-quarters with Southwest presence and route-quarters without, for each carrier.*

Carrier	AA	AS	B6	DL	F9
Difference	−19.1 %	−13.8 %	−6.5 %	−28.4 %	−25.8 %
Carrier	FL	NK	SY	UA	US
Difference	−0.7 %	−33.9 %	1.7 %	−26.6 %	−24.5 %

Tab. 5 shows that almost all carriers charge significantly lower fares if Southwest is present, except for Sun Country (SY) and AirTran (FL). When making a distinction between low-cost carriers and legacy carriers, the sensible result follows that, with Southwest present, legacy carriers decrease their fares by approximately 22 %, while low-cost carriers cut fares by only 13 %.

To summarize, the answer to the research question is:

- On some routes a fare reduction as high as 37.2% is observed after Southwest entry
- Carriers react differently to Southwest presence: legacy carriers reduce fares by approximately 22%, while for low-cost carriers this is only 13%

6 Herfindahl Index and Its Effect on Fares

In this chapter the Herfindahl index will be explained. Also, the effect of the Herfindahl index on the fares will be analyzed. The research question answered in this chapter is: 'How does competition affect fares set by airlines?'

6.1 Herfindahl index

The Herfindahl index (sometimes referred to as the Herfindahl-Hirschmann index) is a term which originates from the field of microeconomics, and is a measurement used to analyse competition within an industry. It is defined as the sum of the squares of the market shares of each firm in that particular industry. Mathematically, it is defined as follows:

$$H = \sum_{i=1}^{n} s_i^2. \tag{8}$$

In Eq. 8, s_i is the market share of each firm expressed as a fraction. As can be seen from the equation, the Herfindahl index can take values between 0 and 1 only. A low number indicates that there are many firms with small market shares, whereas a number closer to one indicates a (near-)monopoly. It is interesting to use the Herfindahl index in this research, because competition is a key factor for airlines when determining their fares.

An interesting way to relate the Herfindahl index to our data, is to look at the Herfindahl index distribution for large and small airports. The results of this analysis are shown below:

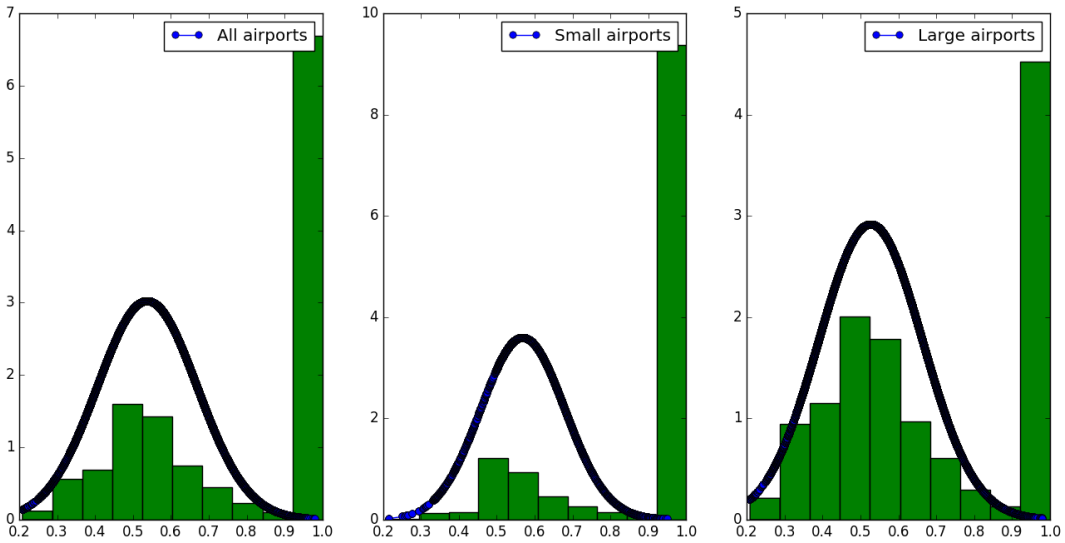

Figure 4: *Herfindahl index distribution for the airports.*

It is interesting to note the peaks around Herfindahl values of 0.5 and 1.0. Peaks around 0.5 indicate that a duopoly is present, and the peak at 1.0 indicate the presence of a monopoly. The curve that can be seen is a curve of the normal distribution, not taking into account the peak at 1.0. It is clear that for both small and large airports the Herfindahl index follows a somewhat similar distribution.

Since a value of 1.0 explains a lot about the competition on airports, another interesting thing to look at would be the number of monopolies for both large and small airports. After performing this analysis, it was found that for

large airports, around 35.5% of the airports have a Herfindahl index of 1.0, whereas this number is 73.3% for small airports. This is quite intuitive, because it indicates that there are many airlines acting on large airports, whereas smaller airports are generally occupied only by low-cost airlines.

6.2 Effect of Herfindahl Index on Fares

The Herfindahl index by itself has shown to be quite an interesting statistic to look at, but it would also be valuable if it is possible to verify the previously found relation between competition and fares.

Airlines that are operating in a monopoly have more freedom when defining their prices. We may therefore expect higher fares on routes with less competition (or a high Herfindahl index), and vice versa for routes with more competition. To perform this analysis, it was important to make a distinction between high and low Herfindahl indices. The results of this analysis are shown in Tab. 6.

Table 6: *Fare/distance for different thresholds of Herfindahl index.*

Hhi Thresholds	Low	Medium	High
0.3/0.8	0.28	0.43	0.49
0.4/0.6	0.31	0.44	0.47
0.45/0.55	0.31	0.45	0.46

As can be seen from Tab. 6, the threshold values to determine in which category a Herfindahl index falls (low or high), were taken at an extreme case first (0.3 and 0.8), and then adjusted (to 0.45 and 0.55). This was done to ensure that the results that are observed are actually trustworthy. One should expect that while the threshold values grow closer together, the difference between the low and high fares also grows closer, but there should still be a notable difference in fares.

This is exactly what can be observed. Initially, it can be seen that the fares are $ 0.28 per mile for the low Herfindahl indices, and $0.49 per mile for the high ones. Finally, the fares grow closer to $0.31 per mile and $0.46 per mile for the low and high Herfindahl values respectively. These results give a clear answer to the research question. Routes with less competition tend to be more expensive than those where more competition is present.

7 Simulated Networks Versus Real Networks

Can mathematical model network simulations be used to simulate ‘real world’ networks? It is difficult to perfectly simulate the evolution of a real world airline network. There are (and have been) factors in play in the development of these networks over time that cannot be reproduced perfectly for every single airline. For instance, the network of a low cost carrier will develop very differently to that of a Low Cost Carrier. This is due to the different business strategies that these 2 airline types use. Also, a ‘legacy carrier’ that has experienced the deregulation of 1978 (like American Airlines) will have developed a different network to Virgin America, which has only been around since 2004.

Therefore the objective of using simulated network models is to see how they compare to real airline networks. The approach used in this chapter can be seen in Fig. 5.

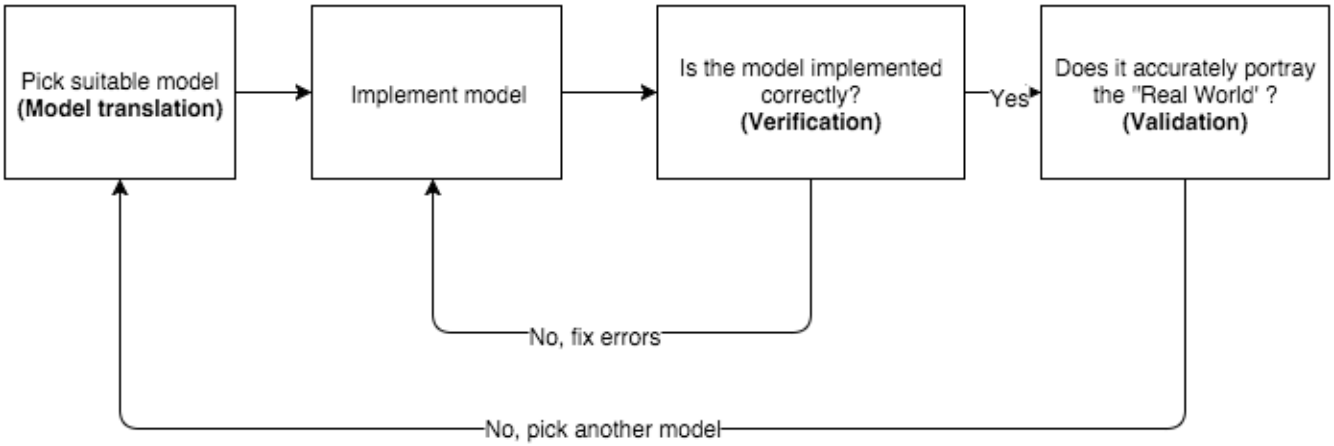

Figure 5: *Flowchart for the implementation of network simulation models.*

Firstly, a suitable network simulation model will be chosen. In the next step, the model will be implemented. For this, the Python programming language and the open source library of scientific tools SciPy were used. After implementing the model, debugging and diagnostic tests are run to make sure the model is implemented correctly (**Verification**). If the model is implemented correctly then the network measures of the simulated network will be compared to an actual 'real world' airline network. In this chapter the Erdős-Rényi model and the Barabási–Albert Model are implemented and studied.

7.1 The Erdős-Rényi Model [5]

There are two closely intertwined mathematical definitions of the Erdős-Rényi model. The first definition is $G(n,p)$, where n is again the number of nodes and p is the independent probability that an edge is present. Therefore the value p can only be between 0 and 1.

The second one specifies a model $G(n,M)$ where n is the number of nodes and M is the number of edges. According to this definition all networks are chosen from a collection of networks which have n number of nodes and M number of edges. In our simulation this definition will be used; n is set equal to the amount of airports Southwest utilizes (88) and a value of 522 is used for M. This is the number of routes which are present in the Southwest network in 2013Q4.

The simulation was run and the network characteristics of the simulated and a real network were analyzed. An explanation of the network characteristics can be found in section 3. A visualization of the two networks can be seen in Fig. 6. It can be seen that the random network looks indeed more 'random' than the real network. In order to better understand this network the characteristics must be examined. These can be found in Tab. 7. In this table it can be seen that the density is almost the same. First off, it can be concluded that the betweenness centrality is a lot higher in the Southwest network. This can be explained by the fact that Southwest uses some degree of 'hubness' in their networks. Therefore the hub nodes will have a relatively high value here. Also, the closeness centrality is significantly higher, which is another sign that points to a hub-and-spoke network. Lastly, the degree centrality is also significantly higher which is another sign of a hub and spoke network. It can be concluded from these network measurements that a real network is not anything like the Erdős-Rényi model. Therefore in the next subsection another model will be discussed which is more suitable.

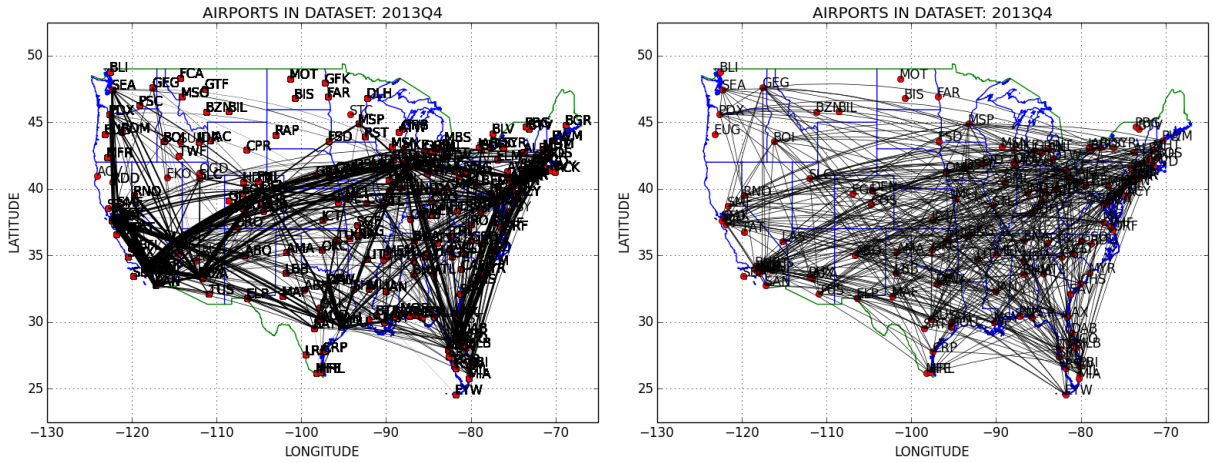

Figure 6: *Visualization of a real network (Southwest airlines 2013Q3) on the left versus the Erdős-Rényi network with $p = 0.6$ on the right.*

Table 7: *Network characteristics of the Erdős-Rényi network and the Southwest network.*

	# nodes	# edges	diameter	density	BC	CC	DC	EC
WN	88	522	3	0.14	0.12	0.73	0.62	0.27
E-R	88	522	4	0.13	0.01	0.50	0.17	0.10

7.2 The Barabási–Albert Model [6]

The Barabási–Albert model is a scale free network model which utilizes preferential attachment. Scale free means that the degree distribution follows a power law. This means that the degree distribution follows a power law. This can be represented as:

$$P(d) \sim d^{-\gamma}. \tag{9}$$

For the Barabási–Albert model γ is equal to three. This model starts with m_0 initial connected nodes. In every iteration a new node is added to the network. The probability that this new node has a connection with node i is as follows:

$$p_i = \frac{k_i}{\sum_j k_j}, \tag{10}$$

where $\sum_j k_j$ denote all the connections in the network. The result of this mechanism is that nodes with a lot of connections tend to accumulate even more connections. Therefore, this model tends to generates hubs inside the network.

This model was also implemented using Python and Scipy. The implementation of this model was verified by checking if the degree distribution was similar to Eq. 9 where γ is equal to 3. The network characteristics of the simulated network compared to that of the real network can be found in Tab. 8.

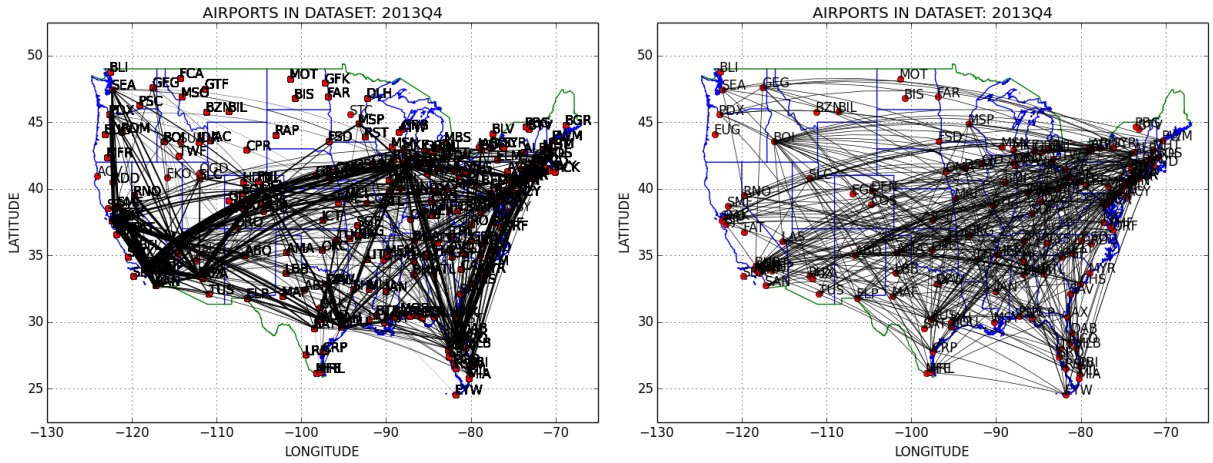

Figure 7: *Visualization of a real network (Southwest airlines 2013Q3) on the left versus the Barabási–Albert network on the right.*

Table 8: *Network characteristics of the Barabási–Albert Model network and the Southwest network.*

	# nodes	# edges	diameter	density	BC	CC	DC	EC
WN	88	522	3	0.14	0.12	0.73	0.62	0.27
B-A	88	507	3	0.13	0.02	0.51	0.15	0.10

It can be seen from Fig. 7 that the simulated network shows more 'hubness' than the Erdős-Rényi simulation. This is can be confirmed when looking at the computed network statistics. The betweenness centrality is higher and the closeness centrality are both higher than those from the Erdős-Rényi simulation.

8 Machine Learning

Can machine learning be used to make predictions on classification problems? In order to answer this question the topic and limitations of this field of study must be explored. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data [7]. In order to do this we split the data set into two parts. A 'training' data set and a 'test' data set. After the algorithm is trained trained the test data will be used as input. The algorithm will then make a prediction on this test data set. The results will then be compared to the actual result to see how this performs. Machine learning can be used to make predictions on either classification or regression problems. An example of a classification problem can be to predict whether an email is spam or not. The prediction value can either be 0 (no spam) or 1 (spam). An example of a regression problem is housing prices. In this example the prediction values can take on any positive real number.

There are 3 different types of 'training processes' in machine learning:

- **Supervised learning:** The algorithm is given the desired outputs during training. In other words, it is being told what is right and wrong
- **Unsupervised learning:** The desired outputs are not given to the algorithm. Instead, it must cluster the data together in order to recognize patterns. It can then make predictions to which group of data points a new data point belongs.

- **Reinforcement learning:** A specific goal is set for this algorithm but it is not taught if it is getting close to its goal of if it is doing things right or wrong.

In this report supervised learning will be used in order to perform predictions on classification problems. The data set will be used in order to make predictions on whether or not Southwest Airlines will be present (1) or not (0) on a ro.

8.1 Machine Learning Theory: Linear Regression [8]

In order to understand how logistic regression is applied, first linear regression must be examined. Below is an example data set containing housing prices:

Table 9: *Example data set.*

Size in feet^2 x)	Price ($) in 1000's (y)
2104	460
1416	315
852	178
...	..

In order to make a prediction on the housing prices if a new data point with only the house size is given a hypothesis must be made. This hypothesis will be in the form of:

$$h_\theta(x) = \theta_0 + \theta_1 x = \theta^T x. \tag{11}$$

Now the parameters θ_0 and θ_1 must be chosen in such a way that they best fit the training data. The distance between hypothesis function and the data points can be calculated by the cost function:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2, \tag{12}$$

where m is equal to the number of datapoints. Now the goal is to choose parameters θ_0 and θ_1 such that the cost function will be minimized and the hypothesis functions best fits the data. The cost function can be either calculated explicitly by setting the derivative to zero or it can be approximated very rapidly by an algorithm such as gradient descent. With gradient descent the algorithm every iteration the algorithm calculates a new θ as such:

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1), \tag{13}$$

where α is the learning rate or step size. In practice, solving such a problem explicitly proves to be very computationally expensive. Therefore gradient descent or another algorithm is used normally used to approximate the minimum of the cost function.

8.2 Machine learning theory: Logistic regression [9]

Logistic regression is used for classification problems where the predicted value will be either 0 or 1. The hypothesis function for logistic regression is slightly different:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}, \tag{14}$$

where $g(\theta^T x)$ is the sigmoid function and where $\theta^T x$ denotes the result (scalar) of the multiplication of the vectors θ^T and x. The sigmoid function has its asymptotes at zero and one. Therefore the sigmoid function always outputs a value between zero and one. This means that the sigmoid function returns the probability $y = 1$ on input x. If $h_\theta(x) = 0.7$ it means that there is a 70% chance of y being 1 and a 30% chance of y being 0. What also follows from this is the fact that if $\theta^T x$ is larger than zero the sigmoid function will be equal or greater than 0.5 which means the system will predict y is equal to 1.

Because the hypothesis function is different the cost function for logistic regression is also different:

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log(h_\theta(x)^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))]. \tag{15}$$

However, again the objective is to minimize this cost function. Therefore the definition of the gradient descent algorithm for logistic regression stays the same and only the definition of the cost function changes.

Another problem which frequently occurs with logistic regression is over-fitting. This means that the hypothesis function tries to fit all the training data points as well as possible. This is illustrated in Fig. 12.

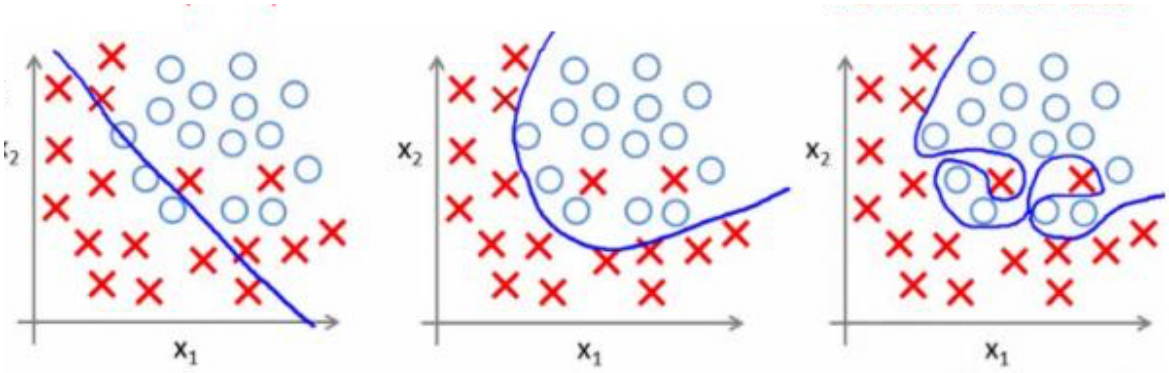

Figure 8: *Illustration of over- and under-fitting of regression and classification problems. [9]*

Because the data is 'over-fitted' the predictions will be less accurate. In order to overcome this a regularization parameter should be added to the cost function as follows:

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log(h_\theta(x)^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2, \tag{16}$$

where λ is the regularization parameter. The result of this new cost function is that now the algorithm also tries to minimize vector θ which means the parameters will be smaller and 'simpler'. λ represents a trade-off between over- and under-fitting and should be chosen carefully.

8.3 Machine Learning Theory: Support Vector Machines (SVM) [10]

In order to minimize the cost function efficiently for logistic regression with a large amount of data points and variables Support Vector Machines are often used. Support Vector Machines use a simplified version of the sigmoid functions which varies linearly instead of exponentially. This will be denoted as $cost_0$ and $cost_1$. Then the cost

function becomes:

$$J(\theta) = C[\sum_{i=1}^{m} y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}))] + \frac{1}{2}\sum_{j=1}^{n} \theta_j^2, \tag{17}$$

where C is the regularization parameter, equal to λ^{-1}. Also the hypothesis function of SVMs does not output a probability as logistic regression does. Instead it outputs 0 or 1. Another advantage of SVMs is that, because of the way the cost function is constructed, it tries to draw the decision boundary with as large a margin possible. This can be seen in Fig. 9.

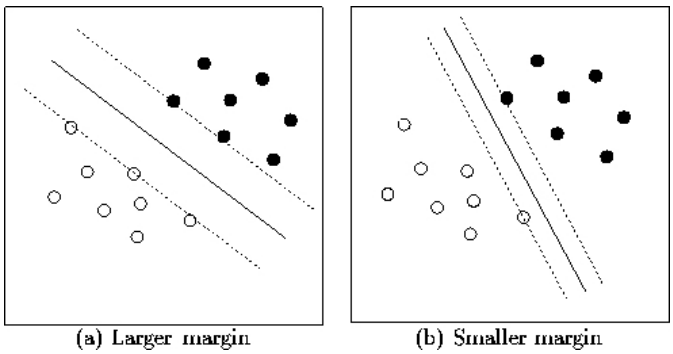

Figure 9: *Large margin classification of Support Vector Machines [10].*

Lastly, SVMs use kernels. This means that the variables in the vector x are replaced by similarity functions:

$$\theta^T f = \theta_0 + \theta_1 f_1 + \theta_2 f_2 + \ldots, \tag{18}$$

where f denotes the similarity function. Explaining kernels mathematically in detail is beyond the scope of this report. What they do, however, is not. Kernels determine the shape of decision boundary's. Different similarity functions lead to different hypothesis functions. Three common kernel functions are the linear kernel (which is the regular hypothesis function) the Radial Basic Function (also know as 'RBF' or Gaussian) kernel and the polynomial kernel. A visual representation of the effect these kernels have on the decision boundary can be seen in Fig. 10.

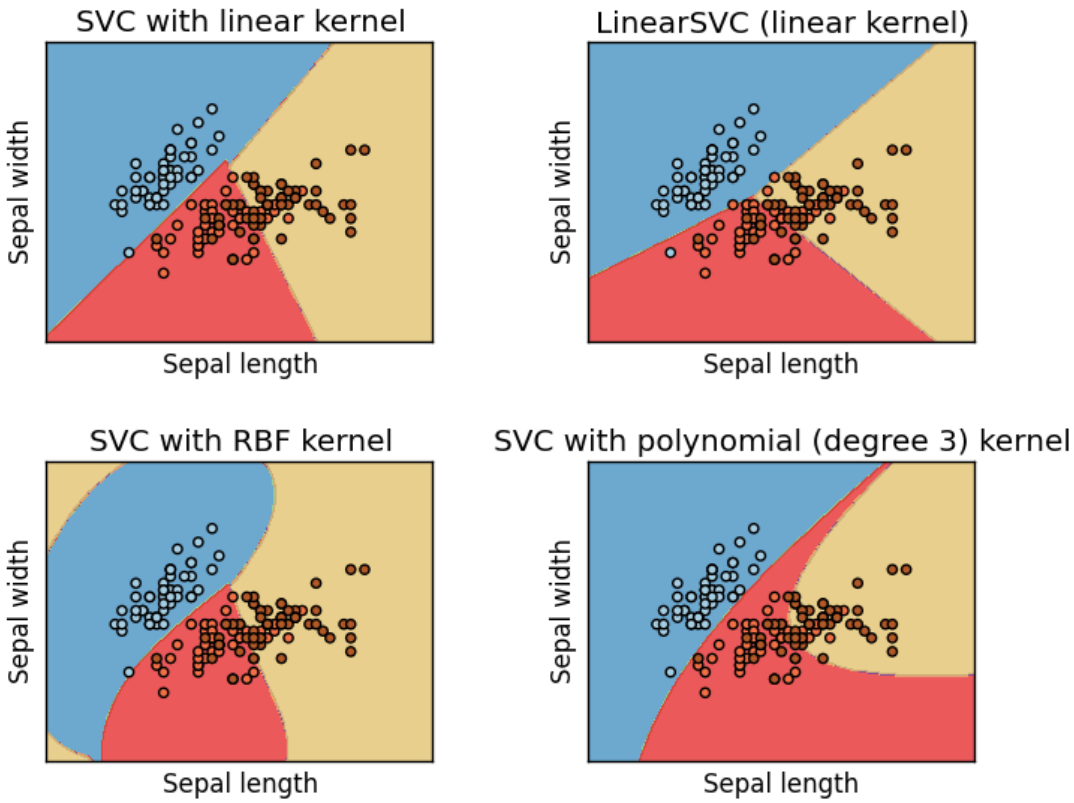

Figure 10: *Illustration of the decision boundary on a one-vs-rest three way classification problem.*

8.4 Applying Machine Learning

The data in the data set is transformed into matrix X with rows i and columns j. The rows represent a route-carrier-quarter and the columns represent the variables or features that are used. All the variables as described in Section 2 about data will be used except for ones concerning the distribution of ticket prices, and the *meanGDPperCapita* variable. The ticket prices will not be used since there are ten variables describing the distribution and this would give a very large weighting to the ticket prices. The *meanGDPperCapita* variable will not be used since this is not available for each data point and will therefore also not be considered. The *WNpresence* classification variable is transformed into a single column vector y. In order to make sure gradient descent is used smoothly all variables should be in between -1 and 1 with a mean around zero. Fortunately, SciKit provides the function 'preprocessing.scale()', which scales the entire matrix X. After the input is scaled the decision boundary or 'support vectors' are calculated. When this is completed, a test matrix X is put into the system and a single column vector y is returned. This column vector contains the predictions on the classification problem.

Firstly, the machine learning problem was run using all variables without the network statistics. Then the network measures were added to the vector X to show how the network variables (as described in Section 2) would affect the prediction. The results can be seen in the figure below:

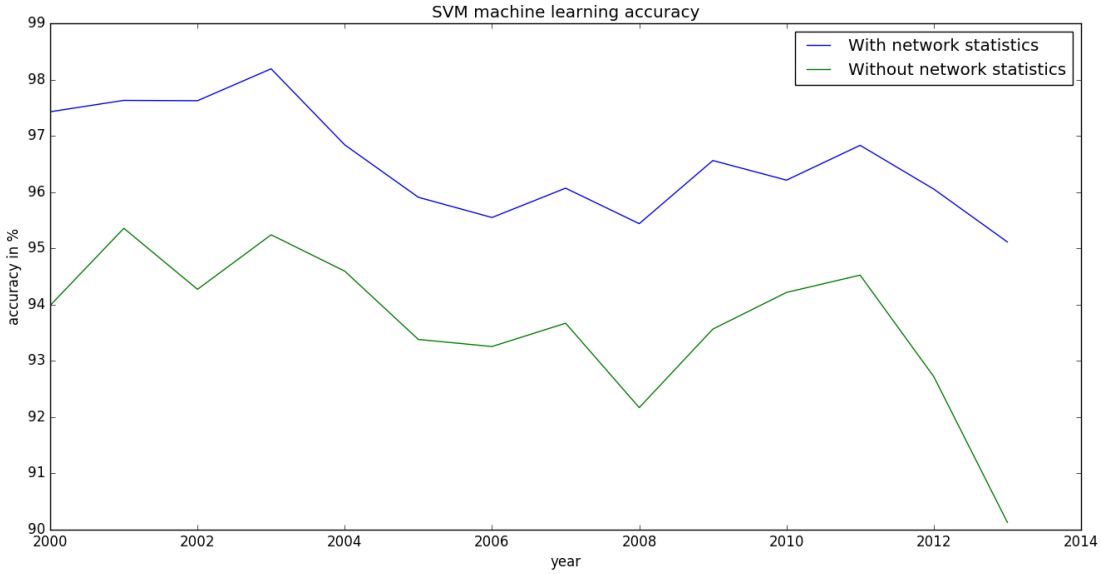

Figure 11: *SVM prediction performance. The blue line depicts the performance of the SVM with network statistics ($C = 1$ $\gamma = 1$).*

The average prediction accuracy of the SVM with network measures is 96.6% compared to 93.6% for the regular SVM. Therefore, the network measures have a profound impact on the performance of the SVM. The prediction becomes much more precise. This simulation was run for 13 years. So thirteen different SVMs have been constructed. For example, if the SVM is trained using the data of 1999 it will make a prediction on the presence of Southwest in the year 2000. The simulation was also run using multiple training years using the network variables but the results proved to be almost the same.This can be seen in the figure below:

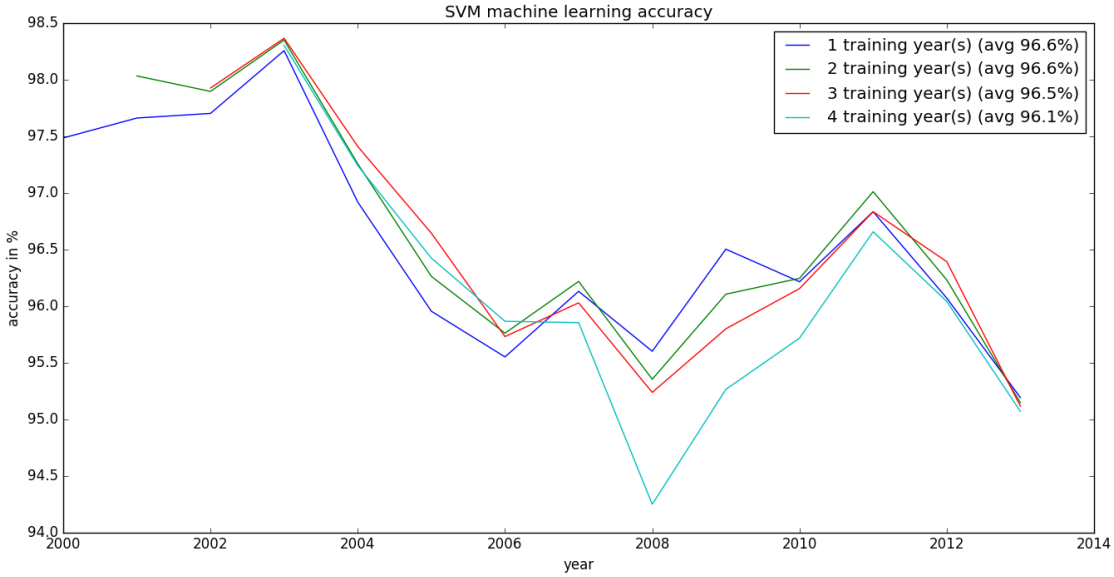

Figure 12: *Performance of SVM using multiple training years.*

This can be explained by the fact that because networks change each year the most recent year will give the best predictions. All of these simulations were run setting the parameter γ for the Gaussian kernel and the regularization parameter to 1.

To conclude, support vector machine learning can give fairly accurate results on this data set for the Southwest presence classification problem. The SVM was observed to function best with a Gaussian kernel. It was seen that, by utilizing network statistic features, the SVMs predictions became more accurate and this tells us that the type of network and network statistics play an important role in whether Southwest will be present or not.

9 Conclusion and Discussion

In this project the U.S Domestic Airline Network was analyzed using data from the DB1B and T-100 databases from the U.S. Department of Transportation. This was performed for a time period of 1999Q1 to 2013Q4. There was an emphasis on analyzing the dynamics of the networks by looking at how statistics change over time.

It was clear that airline networks can be modelled accurately using classical graph theory, on which a number of macro and micro statistics can be computed and interpreted.

The networks in the dataset were modelled in the statistical programming language R in order to gain knowledge about the interaction between fare and other characteristics on a route. It was found that fares are higher on routes between 'richer' cities, on routes with lower load factors, on business routes, and on monopoly and duopoly routes. Also, fares tend to get higher as a route increasingly becomes part of a hub and spoke network.

Additionally, the Southwest Effect was analyzed using the general programming language Python. It was found that the Southwest Effect is indeed present on the U.S Domestic Network, with a fare reduction of up to 37.2% on some routes. It was also found that not all carriers react the same way to the presence of Southwest. Legacy carriers reduce fairs by approximately 27%, whereas for low-cost carriers this reduction is only around 13%.

In order to gain more insight, and to confirm the results found on the effect of competition on fares, the Herfindahl Index was used. Around 35.5% of the large airports have a Herfindahl Index of 1.0, whereas this number is 73.3% for small airports. This shows that there is more competition on larger airports. It was also found that the average fare for routes with a Herfindahl Index of below than 0.45 is \$0.31 per mile, and \$0.46 per mile for routes with a Herfindahl Index larger than 0.55. This clearly emphasizes the result that less competition leads to significantly higher fares.

In order to gain a deeper understanding of the dynamic evolution of networks, the evolution of networks was simulated and compared to the existing networks. The Barabási–Albert model was able to generate hubs and therefore produced network characteristics that were closer to that of real networks. However, different carriers have different characteristics and therefore no single simulation model can be used to simulate all the types of airline networks. Furthermore, there is a certain degree of refinement in real networks that also proves to be difficult to simulate. For instance, the number of nodes in the network of Southwest Airlines is 522 in 2013Q4 while the diameter is only 3.

Finally, machine learning was used in order to predict the classification problem of the Southwest Effect. The Support Vector Machine provided an average accuracy of 96.6% over 13 prediction test sets. Support Vector Machine learning can also be extended to make regression predictions and it will be interesting to see how SVMs will perform on network regression predictions. It is therefore recommended that this should be studied further in the future.

References

[1] D. Robinson and T. Stuart, "Network effects in the governance of strategic alliances," *The Journal of Law, Economics & Organization*, vol. 23, pp. 242–273, 2006.

[2] R. El-Khatib, K. Fogel, and T. Jandik, "CEO network centrality and merger performance," *Journal of Financial Economics*, vol. 116, pp. 349–382, 2015.

[3] Y. Hochberg, A. Ljungqvist, and Y. Lu, "Whom you know matters: Venture capital networks and investment performance," *The Journal of Finance*, vol. 52, pp. 251–302, 2008.

[4] A. Goolsbee and C. Syverson, "How do incumbents respond to the threat of entry? Evidence from the major airlines," *Quarterly Journal of Economics*, vol. 123, pp. 1611–1633, 2008.

[5] P. Erdős and A. Rényi, *On Random Graphs I.* 1959.

[6] A. Barabási and R. Albert, *Statistical mechanics of complex networks*. 2002.

[7] R. Kohavi and F. Provost, "Guest editors' introduction: On applied research in machine learning," *Machine Learning*, vol. 30, pp. 127–132, 1998.

[8] A. Ng, "Machine learning: Week 2, linear regression with multiple variables," Stanford University, Coursera, 2016. Accessed: 3 January 2016.

[9] A. Ng, "Machine learning: Week 3, logistic regression," Stanford University, Coursera, 2016. Accessed: 6 January 2016.

[10] A. Ng, "Machine learning: Week 7, support vector machines," Stanford University, Coursera, 2016. Accessed: 7 January 2016.