BIA-658
SOCIAL NETWORK ANALYSIS
AIRLINE NETWORKS
SHORTEST PATH – CHEAPEST FARE

Ar Krishnasamy Lakshmanaperumal

akrish2@stevens.edu

Problem Statement

- This project is aimed at resolving and providing suggestion to users to select shortest path for the airline route without any connection and would provide suggestion to find the best cheapest price for the route even if it has multiple layover
- The main thing a traveler considers when travelling from Point A to Point B through airplane is time involved in traveling.
- Some travelers doesn't care about time and layovers involved and can travel from Point A to Point B which any layovers involved provided that it is cheaper
- Some travelers doesn't care about cost involved and prefer to reach destination with minimum layover, predominantly business travelers
- To cater the users who care about time we are developing and analyzing a network model to suggest them with shortest time.

Data Set

- In this analysis we have retrieved two datasets one airport.dat file with all the details of airport including ID, Name, City, Country, IATA, ICAO, Lat, Long, Alt, Time zone, DST, Time Zone, Type, and Source.
- And the second dataset routes.dat file contains all airline route and stop details including details like Airline, Airline ID, Source Airport, Source Airport ID, Destination Airport, Destination Airport ID, Codeshare, and Stops Equipment.
- The obtained data doesn't contain price details which if possible, we can find from other sources and can be analyzed for price information but this project we will be focusing on travelling time details only

Data Set – Key Attributes

- The data set obtained has many non-key attributes in it which all can be ignored, and we can consider only key attributes like
  - Source airport,
  - Destination airport, and
  - Number of flights
- We are selecting only these attributes for the following reasons
  - To find and plot a map to visually see the network between source airport and destination airport
  - To identify the number of flight travel from Point A to Point B
  - To inform the user with best path

Data – Observations and Data cleaning

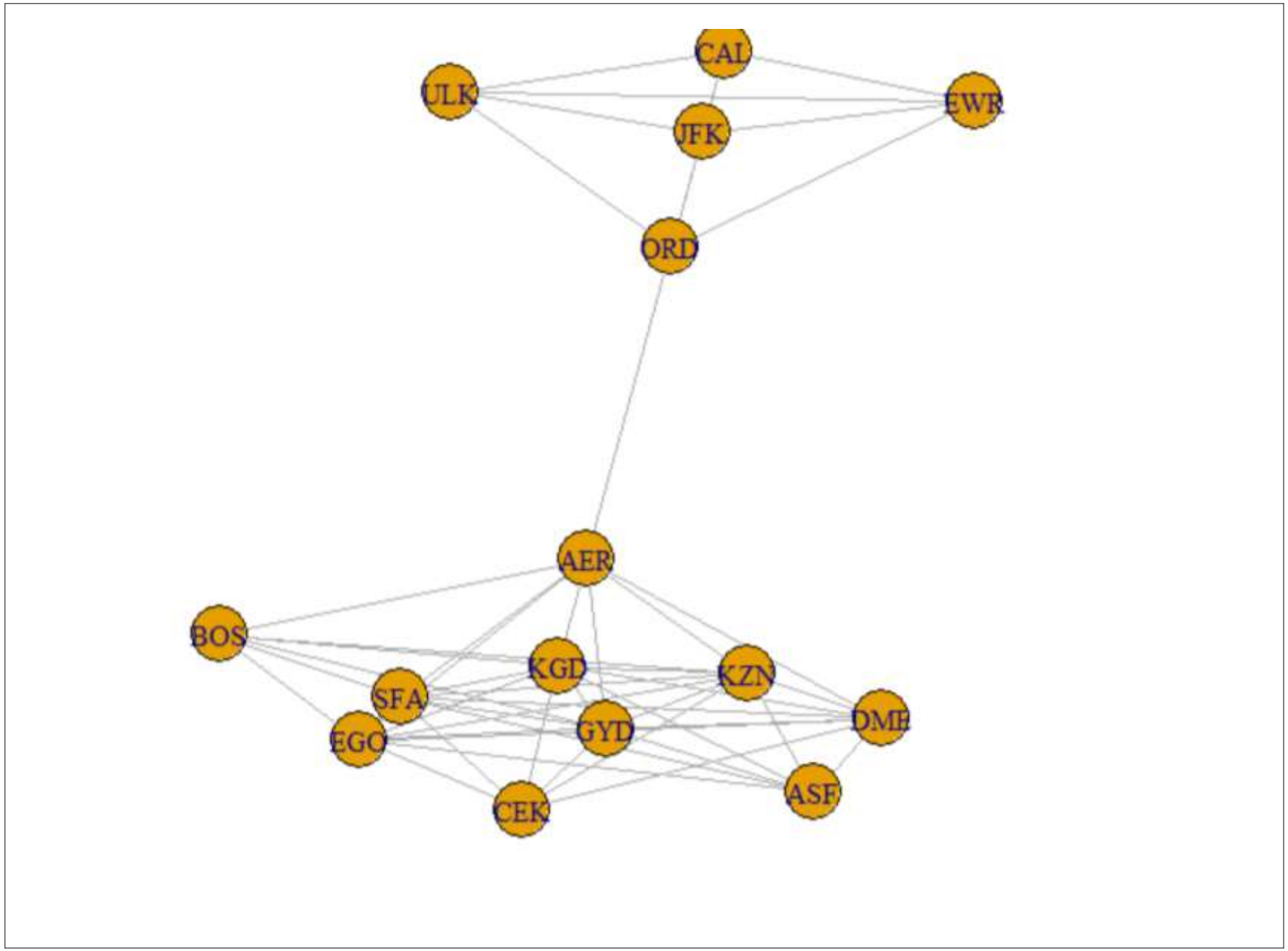
- The dataset obtained has two different files one with airport information's and the other route data set has all routes information with airlines
- We will extract number of flight between airport details by checking any flight with any airlines has routes between source and destination airport and add it to find the count
- Some data obtained will have too many inconsistency like missing data, wrong data format, incorrect-data, out of scope data, duplicate records, irrelevant data and we have followed following methods to address these inconsistency
  - Missing data: the missing data in the final data set is either replaced with dummy value of average number if the attribute is number of flight which is 4 or for fields which didn't have source airport or destination airport, we have removed the entire field from the record.

Data Cleaning

- Wrong data format: some data in source airport and destination airport had different format like entire name instead of 3 letter airport code and since those values are crucial, they are changed to 3 letter code
- In-correct data: some field have incorrect data like the number flight field may contain alphabets instead of number or value like /NA which all can be replaced or removed
- Out-of-scope data: some data like airport which doesn't comes under USA are moved to focus only on flight within USA and numbers which are 0 in number of flights also removed.
- Duplicate records: The duplicate records in the dataset are identified like if the data has same source and destination airport then the records are removed
- Irrelevant data: there might be some records which are not relevant to our analysis are all removed from the analysis

Methods Used

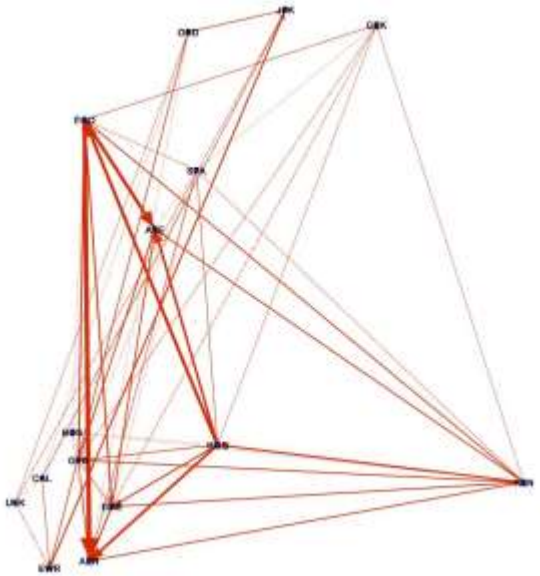
- To find the pattern involved in airline travel we are going to implement a R program to identify the relation between source airport, destination airport and flights involved
- In R program we can the following parameters as dataset:
  - Source = Source Airport
  - Target = Destination Airport
  - Edge = number of Flights
- The network obtained will be too tough to interpolate for any output since we have around 67664 routes entries and 7699 airport details entries
- So, we will reduce the inputs for analysis purpose to 15 routes and airports

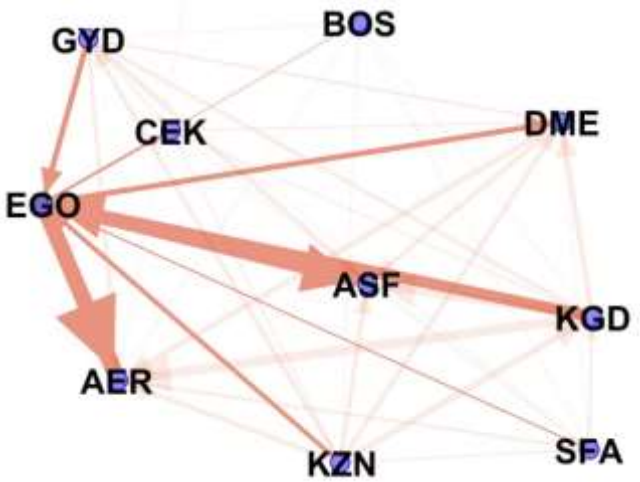


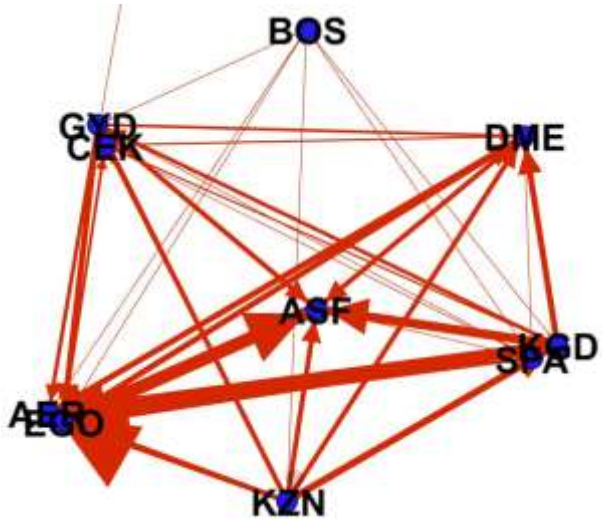
Plot Diagram

- Airline network plot diagram from R programming is plotted to identify the relationship between each airport and flights flying between them

GEPHI - VISUALIZATION







Inferences & Results:

- From the plot diagram it is inferred that the airports are interconnected through various edges either directly or through a middle airport
- Some airlines have direct flights between long distance airports and some airlines take two stops and layover in their hubs to transfer passengers to the next leg of flight
- For example, American airless has JFK as hub, united has EWR as hub and they operate and transfer passengers to smaller airports through their hub
- With this network process we can eliminate the airline which travels through hub/layover and we can select airline which fly's directly to the destination airport
- Many a times this can be costly process, but some sector of people are willing to pay extra for faster travel

Inferences & Results:

- Using R plot diagram, we can visualize the network of airline routes between all the airports and can provide quick relation between source and destination airports
- Also, while further analyzing with airport wise the Gephi mapping diagram is useful in identifying the individual airport route
- For example, in the Gephi diagram 2 it is observed that all flights coming in and out of EGO airport and destination can be identified
- In Gephi diagram 3 we can find the layover airport skipping detail where ASF airport acts as a hub for other airports and some airline doesn't use ASF airport as hub and fly directly

Conclusion & Next Steps

- From this network analysis we have visualized and identified the shortest path between two location in flight
- We can now give suggestion to a person to eliminate layovers and reach destination faster
- Also, with proper pricing data the next step which can be upgraded can be predicting cheaper routes with multiple layovers.
- For example, if the person is willing to spend or spare more time in travelling and layover in airports and expects to pay low fee then the network can be established to figure out multiple airlines together to buy tickets at low price

Data Set Reference

- In this project two datasets are been used their source details are as follows.

- Airline routes dataset:
  - https://raw.githubusercontent.com/jpatokal/openflights/master/data/routes.dat

- Airport details dataset:
  - https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat